Evaluation of exemplar-based generalization and the abstraction of categorical information

Article *in* Journal of Experimental Psychology Learning Memory and Cognition · November 1984

DOI: 10.1037//0278-7393.10.4.638 · Source: PubMed

Jerome R. Busemeyer
Indiana University Bloomington

Doug L Medin
Northwestern University

Journal of Experimental Psychology:
Human Learning and Memory
1981, Vol. 7, No. 6, 418-439

0096-1515/81/0706-0418$00.75

Limitations of Exemplar-Based Generalization and the Abstraction of Categorical Information

Donald Homa, Sharon Sterling, and Lawrence Trepel
Arizona State University

An evaluation of exemplar-based models of generalization was provided for ill-defined categories in a category abstraction paradigm. Subjects initially classified 35 high-level distortions into three categories, defined by 5, 10, and 20 different patterns, followed by a transfer test administered immediately and after 1 wk. The transfer patterns included old, new, prototype, and unrelated exemplars, of which the new patterns were at one of five levels of similarity to a particular training (old) stimulus. In both experiments, increases in category size and old–new similarity facilitated transfer performance. However, the effectiveness of old–new similarity was strongly attenuated by increases in category size and delay of the transfer test. It was concluded that examplar-based generalization may be effective only under conditions of minimal category experience and immediacy of test; with continued category experience, performance on the prototype determines classification accuracy.

Categories are said to be ill defined (Neisser, 1967) when it is not obvious what dimensions characterize a category, and the variety among the potential members of a category is essentially infinite. Examples of ill-defined categories are quite diverse and would include the natural categories, musical style, hand-written letter *A*s, and the class of sound patterns associated with a specific spoken word.

How the human organism learns ill-defined categories, and how this knowledge is transferred to novel situations, has been a topic of considerable attention over the past 10 years. Posner and Keele (1968, 1970) argued that a prototype or central tendency is abstracted during the classification of distorted but related patterns. In their experiments, the subject initially sorted dot-pattern stimuli into a number of categories, with each category represented by a different reference pattern (objective prototype). Classification of old, new, and prototype patterns on a transfer test, administered immediately and again after a delay of 1 week, revealed that significant forgetting occurred for the old training stimuli. However, classification of the objective prototype was unaffected by this delay period. As a consequence, Posner and Keele (1970) argued that the objective prototype could not have been classified on the transfer test via generalization to the old, stored patterns, since any performance decrement on the old patterns should have been accompanied by a similar decrement for the prototype. Rather, they proposed that the subject abstracted the prototype during the classification phase and that the prototype was simply resistent to decay. The stability of prototypical performance, within the context of a deterioration for the old patterns, has now been obtained in numerous experiments (Homa, Cross, Cornell, Goldman, & Schwartz, 1973; Homa & Vosburgh, 1976; Strange, Kenney, Kessel, & Jenkins, 1970), in which delays varied from 4 days to 10 weeks.

These results do not suggest, however, that old information is unavailable or lost completely from storage. In fact, most experiments have found that performance on the old patterns is superior to performance on the new instances, even after lengthy time

This research was supported by National Institute of Mental Health Grant 5 R01 MH28270.

The authors would like to thank Julie Omohundro and Harold Makoany for their comments on an earlier draft of this manuscript. Portions of this study were presented to the Western Psychological Association, San Diego, April 1979.

Requests for reprints should be sent to Donald Homa, Department of Psychology, Arizona State University, Tempe, Arizona 85281.

delays. For example, in the experiment by Homa and Vosburgh (1976), performance accuracy on the old learning stimuli exceeded that of the new patterns at the same level of distortion by 10% to 20% after 10 weeks. As a consequence, it seems likely that specific information, as well as information about the central tendency of the category, remains in memory. However, transfer to novel stimuli is thought to be accomplished primarily via generalization to the abstracted prototype, especially after lengthy time delays. Support for this view has also been based on transfer performance; with variable time delays, transfer performance for novel patterns is also unaffected, although some deterioration may occur for high-level distortions of a category (Homa & Vosburgh, 1976).

Recently, an alternative view of category learning has been proposed that is based on generalization to the stored exemplars (Brooks, 1978; Medin & Schaffer, 1978). According to this view, classification of a novel stimulus is determined by its similarity to the stored exemplars. The differential retention that is typically found for old and new stimuli (e.g., Posner & Keele, 1970) is accounted for, according to one version of exemplar theory (Medin & Schaffer, 1978), by alterations of the dimensional weights that characterize the stimuli. Although this prediction has not received empirical confirmation, Hintzman and Ludlam (1980) were able to simulate the differential forgetting of prototype and old exemplars with a computer model that used the stored exemplars as the sole basis for generalization. Here, forgetting of stimulus properties occurred in an all-or-none manner, and only a single exemplar was retrieved from each category at the time of transfer. Nonetheless, a number of procedural differences exist between these experiments and those which have used artificial categories: (a) The categories used by Brooks (1978) and Medin and Schaffer (1978) were not ill defined; (b) category experience is usually minimal, in that as few as three to six stimuli may define a category in the learning phase; and (c) learning variables known to shape generalization gradients in the category abstraction paradigm (e.g., category size, stimulus distortion, and category similarity) were not manipulated in these experiments.

For example, in the experiments by Medin and Schaffer (1978), the stimuli consisted of forms that varied along four binary-valued dimensions (such as red–green; triangle–circle), only two categories were learned,[1] and as few as three patterns may have represented each category. Each of these characteristics may be contrasted with research that has used ill-defined stimuli (e.g., Homa, 1978); the stimuli are typically statistically distorted forms, where the potential size of the population is essentially infinite, the dimensions underlying each category are obscure, numerous categories may be learned, and as many as 30 different patterns may represent a category during learning. As a result of these differences in stimulus composition, it makes sense to manipulate stimulus variables when ill-defined categories are investigated, whereas variable manipulations are largely precluded when well-defined stimuli are used. For example, the size of the stimulus population in the Medin and Schaffer (1978) study was 16, a number that is further reduced by using more than one category and by using different stimuli in the learning and transfer phases. Given these constraints, manipulations of variables like category size in the learning phase become virtually impossible.[2]

[1] An interpretative problem with the experiments by Medin and Schaffer (1978) is that subjects were not taken to an errorless learning criterion before the transfer test was given; in Experiments 2–4, more than 50% of the subjects failed to reach criterion. As a consequence, it is questionable whether generalization to old patterns can be properly evaluated, given that there is no assurance that the old patterns were stored in memory.

[2] These same concerns may be registered about the computer simulation by Hintzman and Ludlam (1980), who held constant the number (two) and similarity (unknown) of categories during learning, the degree of pattern distortion during learning and transfer (low), and so on. It would be worthwhile to know whether the results simulated in their study are specific to the conditions explored, or whether their model can be extended to other situations as well, such as the use of two or more categories, or variable levels of pattern distortion in learning and transfer. Also, stable classification of the category prototype has been obtained under conditions in which a default or "none" category is available at the time of transfer (e.g., Hartley & Homa, 1981; Homa & Hibbs, 1978). Had a default category been

The importance of learning variables on category abstraction when ill-defined categories are used has been inferred from the effect of variable manipulations during learning on subsequent transfer to novel patterns. For example, when the number of patterns that defines a category in learning is increased, later transfer to novel patterns on an immediate (Homa, 1978; Homa et al., 1973; Homa & Vosburgh, 1976) and delayed (Homa et al., 1973; Homa & Vosburgh, 1976) test is enhanced. Furthermore, when the number of categories discriminated during learning is increased (Homa & Chambliss, 1975) or when the within-category stimulus variance is increased for categories defined by numerous exemplars (Homa & Vosburgh, 1976), generalization to new patterns is improved. In these studies, it has been argued that the abstracted prototype evolves with or is modified by exemplar experience, in which the aforementioned learning variables play a critical role in shaping the subject's knowledge of a category and its breadth.

A second reason for considering the role of learning variables on category abstraction is that they are an inevitable by-product of most naturally occurring learning situations. For example, the human organism most likely acquires information about concepts by exemplar experience. Not only are the examples of most naturally occurring categories likely to span a wide range of variation, but the sheer number of exemplars of a category is countless. As such, evaluations of categorization models that are based on as few as three examples of a category (e.g., Medin & Schaffer, 1978) may be misleading. Nonetheless, exemplar-based models of classification have not been directly evaluated in a category abstraction paradigm, in which the categories are ill defined.

In the present study, the effectiveness of old–new similarity on subsequent transfer to novel stimuli was determined for five levels of similarity. Thus, each new pattern presented on the transfer test had a specific similarity relationship to a given old training stimulus. Otherwise, each new pattern was a high-level distortion of the category prototype and, therefore, roughly equidistant from the prototype. In addition, old–new similarity was evaluated in the context of two variables known to influence subsequent generalization performance: category size and the time of the transfer test.

Predictions of Exemplar and Prototype Models

In general, exemplar-based models of classification predict that performance on a transfer test will be heavily influenced by the degree of similarity of a new pattern to a particular old stimulus. Three different versions of exemplar-based generalization were considered in the present study, a single member (SM) model, a fixed set (FS) model, and a complete set (CS) model. These models differ only in the number of stored exemplars retrieved in response to the presentation of a test stimulus at the time of transfer. According to the SM model, the stored member that is most similar to the presented stimulus is retrieved from each category, and classification occurs to that category that results in the best similarity match (Hintzman & Ludlam, 1980). In the FS and CS models, the subject retrieves either a fixed number of stored members from each category or the entire set of stored patterns, respectively, at the time of classification. For example, if three categories were each defined by 20 different exemplars, then the number of patterns retrieved per category would be 1, N, and 20 for the SM, FM, and CS models, respectively, where N is

available in the Hintzman and Ludlam simulation, it is less clear whether stable performance would have been obtained for the new and prototype stimuli. Specifically, since all stored items underwent progressive deterioration in the simulation, it seems inevitable that an increasing percentage of old, new, and prototype stimuli would be assigned to the default category. Finally, the simulation was conducted on well-defined patterns in which the total pool of acceptable properties were specified at the outset. From introspective reports, as well as recent pilot work of ours, it is likely that the number and quality of features are not constant for ill-defined categories, but may change with the level of learning (a point also noted by Fisher, 1916, and Hull, 1920). If it could be demonstrated that the patterning of results obtained by Hintzman and Ludlam is robust across these conditions, then obviously examplar-based and prototype-based models of classification cannot be distinguished by the differential forgetting of prototype and old exemplars.

some number between 1 and 20. Whereas the CS model is similar to the context model of Medin and Schaffer (1978), there has been little investigation of examplar-based models involving the retrieval of a fixed set of stored patterns (Reed, 1972, is an exception). The major motivation for considering the FS model is based on results that suggest that only a portion of the original category members are available at the conclusion of learning (Omohundro 1981; Homa, Omohundro, & Courter, Note 1). For both the SM and FS models, it is assumed that the stored exemplar most similar to the transfer stimulus is always included in the retrieved set of exemplars.

To clarify the predictions of these models for the present study, let x be the similarity between a test probe and the closest same-category member that is stored, let s_w be the average within-category similarity to the remaining members of the category, and let s_b be the average between-category similarity of the test probe to members of alternative categories. Then the classification algorithm for the CS exemplar model (Medin & Schaffer, 1978) appropriate to the present study would be

$$E_{a,i} = (x + (n_a - 1)s_w)/ \\ (x + (n_a - 1)s_w + n_b s_b), \quad (1)$$

where $E_{a,i}$ = evidence favoring classification of Pattern i into Category A, n_a is the number of patterns stored in Category A, and n_b is the number of patterns stored in alternative categories.

For the SM exemplar model, Equation 1 reduces to

$$E_{a,i} = x/(x + (M - 1)s_b), \quad (2)$$

whereas for the FS exemplar model, Equation 1 can be rewritten as

$$E_{a,i} = (x = (N - 1)s_w)/(x + (N - 1)s_w \\ + (M - 1)Ns_b), \quad (3)$$

where M = number of categories learned, and N = number of stored exemplars retrieved per category at the time of classification.

If we assume that only the similarity between the test probe and the closest within-category member is varied (i.e., x is varied), and that the average within- and between-category similarity is held roughly constant for the remaining patterns, then classification accuracy ($E_{a,i}$) for all three exemplar models should be a monotonically increasing function of old–new similarity. These models do make one differential prediction, however. Both the SM (Equation 2) and FS (Equation 3) exemplar models predict that old–new similarity and the size of the category should be additive. That is, the effect of the similarity between the test probe and its closest stored member on classification performance should not interact with the number of patterns that originally defined the category in learning. In contrast, the CS (Equation 1) exemplar model can predict an interaction between old–new similarity and category size, because the *relative* contribution of specific old–new similarity on classification is diminished by increases in category size.[3]

The predictions of a prototype model on classification vary depending upon how the prototype is characterized and whether information other than the central tendency is stored. In most classificatory problems, three potential sources of information can be identified: (a) the central tendency or abstracted prototype for that category; (b) specific information about the exemplars that defined the category during learning; and (c) the boundary or breadth of the category. If only the prototype is stored follow-

[3] Equation 1 is a general form for a class of exemplar models. However, it should be noted that the similarity between two patterns in the context model (Medin & Schaffer, 1978) is determined in a multiplicative manner for the component dimensions. As such, two stimuli that are highly similar along all dimensions but one could be viewed as quite dissimilar if that disparate dimension was both salient and psychologically discriminating to the subject. In the present study, any two high-level distortions (except for the manipulated transfer pattern and its stored counterpart) should be viewed as quite dissimilar, since the average distance moved per dot between two high-level patterns is quite large. More critically, the likelihood is near certainty that at least one of the corresponding points in the two high-level patterns will be sufficiently separated to reach values obtained by randomly related patterns.

ing the learning phase, then classification of new patterns should be determined primarily by the similarity of the test patterns to the prototype; that is, classification should be monotonically related to the distortion level of the test pattern. Since all transfer patterns in the present study were high-level distortions, a pure prototype model would predict that all new patterns would be classified equally well and independently of old-new similarity.

According to a mixed model, all three sources of information (prototype, specific exemplar information, and category breadth) may be available at the time of transfer, but their availability is modulated by the amount and degree of exemplar experience provided during learning. If few patterns represented a category during learning, the subject's representation for that category may consist primarily of information about the specific exemplars. As the degree of exemplar experience is increased, the representation of the category is increasingly dominated by information about the central tendency and the breadth for that category. The role of specific exemplar information is further reduced by time delays between original learning and test. The major predictions of a mixed prototype model in the present study are that old-new similarity should be an effective variable primarily when the category size is small and the transfer test is given immediately. For categories defined by increasing numbers of exemplars, the importance of old-new similarity should be diminished, especially on a delayed test. In effect, a mixed prototype model predicts an interaction between old-new similarity and time of test. Finally, a mixed prototype model predicts that accurate classification of the objective prototype, relative to other novel patterns, should be a function of variable manipulations that affect category abstraction. In particular, defining a category by few high-level distortions may result in minimal abstraction. As a consequence, the objective prototype may be classified more poorly than new patterns that are highly similar to old patterns. However, classification of the objective prototype should exceed that of novel stimuli when category size is large and the test is delayed, regardless of old-new similarity.

Mixed Prototype Versus Complete Set Exemplar Model

Given that the SM and FS exemplar models predict that old-new similarity and category size should be additive, whereas the CS exemplar model and the mixed prototype model both predict an interaction between these two variables, it may be asked whether the latter two models can be distinguished. In the present study, support for the CS exemplar model would be obtained if the derived values of x (Equation 1) were lawfully related to objective old-new similarity and independent of category size. If the derived values of x were found to be independent of objective old-new similarity and/or interacted with category size, then it would follow that factors other than objective old-new similarity were responsible for classification accuracy. A similar conclusion would result if the derived values of x approached 0.00. In this case, the contribution of specific old-new similarity to classification would be negligible. In summary, the CS exemplar model would be strongly supported if classification performance were shown to be an interactive function of old-new similarity and category size, even though the derived values of x were systematically related to old-new similarity and independent of category size. The mixed prototype model would be supported if significant interactions between old-new similarity and category size and between old-new similarity and time of test were obtained for the classification data, and the derived values of x were independent of objective old-new similarity.

Two additional controls were introduced to maximize the likelihood that exemplar generalization might occur; the transfer test was administered only following errorless classification of the learning stimuli, thus insuring that the transfer patterns were well learned, and all transfer patterns were high-level distortions (Posner, Goldsmith, & Welton, 1967) of a category prototype. The latter control guaranteed the occurrence of transfer patterns that were markedly more similar to a learning pattern than to the prototype.

Experiment 1

In Experiment 1, the transfer phase followed a criterial classification phase in which categories were represented by 5, 10, and 20 different high-level distortions of a prototype. High-level distortions share little obvious physical similarity to each other (Homa, Rhoads, & Chambliss, 1979). In terms of physical distortion, the high-level distortion stands roughly midway between the category prototype and unrelated patterns (Posner et al., 1967).

On the transfer test, administered both immediately after learning and after a week's delay, the subject again classified—without feedback—old, new, prototypical, and unrelated stimuli. Of major interest was the similarity relationship between the old and new stimuli on the transfer test. Each new stimulus had a manipulated similarity (distance) relationship to one of the original learning stimuli. Specifically, 20% of all new patterns were minimal distortions of a particular learning stimulus, having an average distance from an old pattern of 1.00 Euclidean unit/dot. The remaining new patterns were either 2.00 units/dot (20%), 3.00 units/dot (20%), 4.00 units/dot (20%), or 5.00 units/dot (20%) from a particular old stimulus. All new patterns, regardless of the old–new similarity relationship, were high-level distortions of a prototype. That is, all new patterns were, on the average, about 4.60 units/dot from the prototype. Thus, all new patterns on the transfer test were closer or more similar to a particular old training stimulus than to the category prototype, except for those new patterns which were 5.00 units/dot from an old pattern. Figure 1 shows an example of a prototypical form, a particular old stimulus, and five new transfer stimuli, each at a different level of old–new similarity from the old stimulus.

Two comments regarding the construction of the new stimuli on the transfer test should be mentioned. First, each vertex of an old pattern was displaced by an *equal* amount in the construction of a new pattern at a particular old–new similarity level. In effect, the distortion of a new pattern from a particular old stimulus was uniformly applied to each vertex. Second, each new pattern was always closer to its designated old pattern than to any other training stimulus, even when the old–new similarity value was 5.00 units/dot (a high-level distortion of an old pattern). The sampling distribution for distances between pairs of high-level distortions typically varies from about 5.00 to 9.00 units/dot (Homa, 1978). To verify this for the stimuli employed in the present study, the average distance moved per dot was computed for all old/new patterns for each prototype; a subset of the old–new similarity matrix is shown in Table 1. For example, Transfer Pattern 1 was a new pattern that was 5.0 units/dot from the prototype, 2.0 units/dot from its designated old pattern (Pattern 11), and 6.0–8.9 units/dot from the remaining 19 old patterns. Similarly, Transfer Pattern 4 was a new pattern that was 5.1 units/dot from the same prototype, 5.0 units/dot from its designated old pattern (Pattern 14), and 5.3–8.6 units/dot from the remaining patterns.

Method

Subjects. A total of 24 Arizona State University undergraduates served as subjects. Two subjects were replaced, due to abnormally high error rates on the transfer test.

Materials and apparatus. Members of three form categories (P2, P3, P4) served as stimuli. Construction of these forms has been described previously (Homa, 1978). Briefly, a form category is created by first generating a random nine-dot pattern, and then connecting the dots with lines. This pattern is arbitrarily designated as a prototype. Different members of this category are then generated by statistically moving each of the dots of the prototype. Generally, patterns that are high-level distortions share little obvious similarity to the prototype, and low-level distortions appear very similar to the prototype. The dots of each pattern occupy locations within a 50 × 50 unit grid. For high-level distortions, each dot is displaced, on the average, about 4.6 units from each corresponding dot of the prototype. Thus, the topography of a category can be thought of as a hypersphere with the prototype located in the center and the high-level distortions on the surface of the sphere. The radius of the sphere would be equal to 4.6 units, and the average chord distance between any two high-level distortions on this sphere would be about 7.5 units. The new transfer stimuli would also reside on the surface

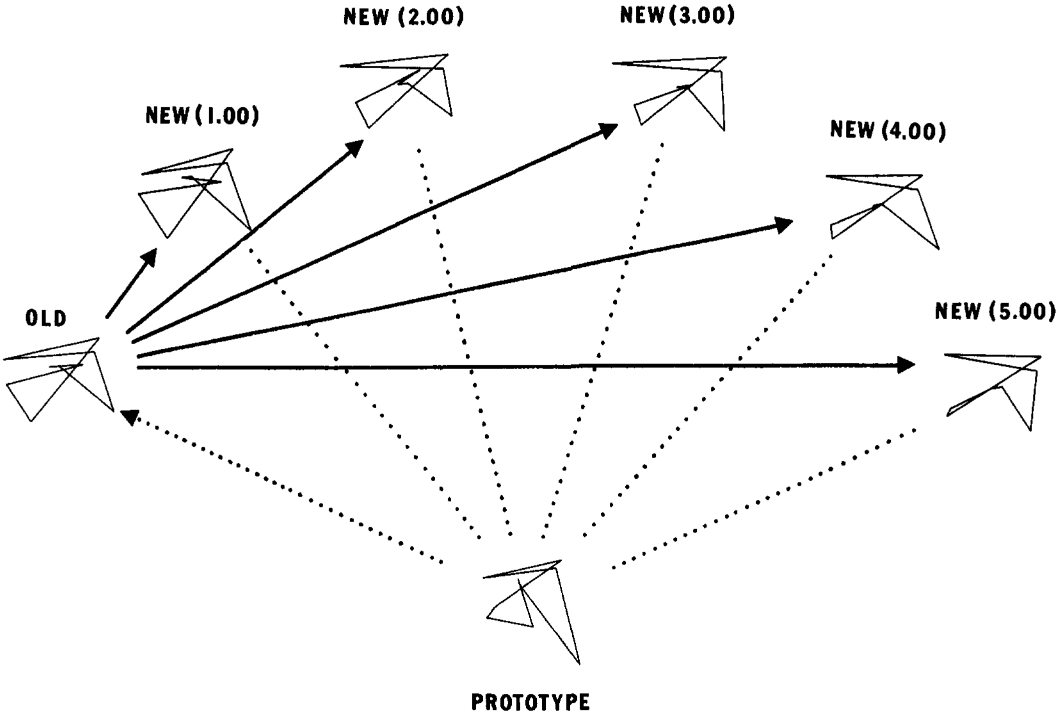

Figure 1. Example of a prototypical form, a high-level training (old) distortion, and five transfer (new) patterns. (The five transfer patterns are equidistant from the prototype and at one of five distances from the old pattern.)

of the sphere, but would have chord distances to a particular old pattern of 1.0–5.0 units.

Unlike previous studies that manipulated the degree of distortion of a pattern to the prototype, the present study systematically varied the degree of distortion of a new pattern to a particular old pattern. Each dot of a new pattern was displaced exactly 1.0, 2.0, 3.0, 4.0, or 5.0 units from the dots in a particular old pattern. The degree of distortion of the new pattern was then computed relative to the prototype; if the degree of distortion from the prototype was either too large or too small, the direction (but not the magnitude) of the dot movement of the new pattern was adjusted until the new pattern was both a high-level distortion of the prototype and at its proper distance from an old pattern.

The basic stimulus pool consisted of 228 patterns. Of these, 71 belonged to each of three different prototype categories, and 15 functioned as foils in the transfer phase; these latter 15 stimuli were essentially random patterns that were statistically unrelated to the three prototype categories.[4] For a given category, the 71 patterns consisted of the category prototype, 20 training patterns (old), and 50 high-level transfer patterns (new). These patterns were drawn by a Cal-Comp plotter and affixed to 6 × 9 in. cards. During the learning and transfer phase, each stimulus was hand held by the experimenter, with the subject viewing the stimulus from across a small table.

Procedure. The subject was told that a series of patterns would be shown in which the task was to determine which patterns belonged to the same category. The subject was instructed to classify the forms into three groups, called A, B, and C, and told not to expect an equal number of patterns in each group. Stimuli were

[4] It is not obvious how one should measure the distance between two patterns from different categories, since corresponding points in two patterns from different categories cannot be determined. If the correspondence is made randomly, then the average Euclidean distance moved per point is about 15.00 for patterns from different categories. If corresponding points are defined in terms of a best fit, then this distance can be reduced to about 10.00 units/point. This is achieved by computing the average distance for all pattern rotations and defining the best fit as the rotation that minimizes point separation. To date, we have not found that this value (10.00 units/dot) is further reducible. Still, this minimum value is not substantially greater than that obtained between two high-level distortions from the same category (about 7.50 units/dot). Regardless, it seems likely that the objective representation of the categorical space is composed of nonoverlapping spheres, with the distance between spheres (categories) still undetermined.

Table 1
Mean Interpoint Distances Between 10 New and 20 Old Patterns for One Category

New		Old (training) patterns																			
pattern	P	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	5.0	8.1	8.3	8.9	8.3	7.5	6.0	7.4	7.2	6.2	6.0	*2.0*	6.8	8.2	6.0	7.5	8.4	6.9	6.7	7.2	7.8
2	4.8	6.0	6.7	7.3	7.6	7.0	4.4	6.6	6.6	6.3	6.8	6.5	*3.0*	7.3	5.8	7.9	7.0	5.6	7.6	8.1	7.4
3	5.3	7.7	7.4	8.9	6.1	7.4	6.2	8.0	5.7	6.8	6.5	6.5	7.7	*4.0*	7.1	7.2	6.3	6.7	7.9	5.5	8.4
4	5.1	8.5	7.1	8.6	8.4	7.4	5.3	6.2	6.2	6.2	5.8	5.8	6.9	7.1	*5.0*	8.3	8.0	7.2	7.3	7.3	8.0
5	4.8	6.5	7.6	7.6	4.8	7.7	7.5	7.4	7.4	7.1	6.8	6.3	7.2	7.0	7.3	*1.0*	7.1	6.8	7.6	7.3	7.7
6	5.2	6.9	7.1	7.8	7.1	7.4	5.0	7.4	6.1	5.5	7.2	7.6	6.4	6.4	7.7	7.9	*2.0*	7.6	8.5	6.9	8.1
7	5.2	6.2	6.9	7.8	7.2	8.5	4.6	7.6	5.9	7.1	6.4	6.4	5.2	6.7	6.1	7.1	7.1	*3.0*	7.8	8.5	8.2
8	5.0	7.4	6.6	8.0	7.6	5.9	6.3	5.4	7.2	6.7	6.7	6.0	6.6	7.7	6.4	7.5	8.7	8.7	*4.0*	6.5	7.7
9	5.2	8.3	7.2	9.4	6.0	7.6	6.0	7.0	6.7	6.8	6.9	6.0	7.8	6.6	7.2	6.4	6.9	7.9	7.8	*5.0*	7.7
10	5.1	7.6	6.6	7.4	7.0	6.9	8.1	8.4	5.7	7.3	6.8	6.7	6.8	7.1	5.5	7.5	7.7	7.0	7.3	7.2	*1.0*

Note. P = prototype. Italicized values refer to manipulated levels of old–new similarity.

presented one at a time, and learning was self-paced. Each response was followed by correct feedback ("no, that is a B pattern"), and learning was ended after two consecutive errorless trials. Each trial contained 35 different stimuli, 5 belonging to one category, 10 to another, and 20 to the third. Four different random orders of the 35 learning stimuli were used to present the stimuli.

The transfer phase began immediately after completion of the learning phase. Each subject was tested twice, once immediately and once after a delay of 1 wk. A total of 111 transfer stimuli were presented, 15 belonging to none of the three learned categories (foils), and 32 from each of the three learned categories. The 32 stimuli that belonged to each learned category consisted of 2 copies of the category prototype, 5 old (learning) patterns, and 25 new patterns. The 25 new patterns were distributed evenly across five levels of old/new similarity; that is, 5 patterns were 1.00 unit/dot from a particular old pattern, 5 were 2.00 units/dot from another old pattern, and so on. The 111 transfer patterns were presented in one of three different random orders, and no feedback was provided. Prior to the transfer task, the subject was told, in part: "A small percentage of patterns belong, in fact, to none of the three categories. If you feel that a given pattern doesn't belong to one of the three categories—and about 10% will not—then assign that pattern to a 'junk' category." The subject was also told to assign a confidence value to each classification, 1 indicating little confidence in their assignment, 2 indicating that they were somewhat confident, and 3 indicating that they were very confident of their choice. Finally, each subject was told that, unlike the learning phase, each category would be represented by the same number of patterns.

Design. The major variables of category size (5, 10, 20), time of test (immediate, 1 wk. delay), stimulus type (prototype, old, new, random), and old–new similarity (1.0–5.0 units/dot) were manipulated in a within-subject design. A Greco-Latin square was used to assign stimuli to the factors of category size, prototype representing each category (P2, P3, P4), and name assigned the category during learning (A, B, C). A total of eight subjects were randomly assigned to each row of the resulting square. As a consequence, each category size was represented equally often by each prototype and each category name.

Results

Original learning. The mean number of trials to reach criterion was 16.96, with about 80% of the subjects requiring 12–21 trials. The trial of last error for any pattern belonging to categories represented by 5, 10, and 20 patterns was 12.75, 13.92, and 13.29, respectively ($p > .05$). Thus, speed of learning was not different for categories represented by different numbers of patterns.

Transfer performance. As expected, overall performance systematically improved with increases in category size; for

5-, 10-, and 20-instance categories, accuracy of classification was .678, .753, and .853, respectively, $F(2, 46) = 16.73$ (MS_e = 1,160), $p < .001$.[5] The facilitative effect of increasing category size was especially pronounced for the prototype stimulus, with an accuracy difference of 40% for categories defined by 5 and 20 instances; the magnitude of this facilitation for new and old patterns was about 17% and 7%, respectively. As might be expected, the Category Size × Stimulus type interaction was highly significant, $F(4, 92) = 7.83$ ($MS_e = 592$), $p < .001$.

The main effect of stimulus type (old, new, prototype) was highly significant, $F(2, 46) = 19.04$ ($MS_e = 559$), with old stimuli (.885) classified significantly better than new (.740) or prototypical (.732) stimuli. Although the main effect of delay was not significant ($F < 1$), the interaction between stimulus type and delay was significant, $F(2, 46) = 5.20$ ($MS_e = 188$), $p < .01$. The latter interaction was due to the fact that classification accuracy for old stimuli slightly deteriorated across the week delay (.906 vs. .864), whereas performance on the new (.732 vs. .747) and prototype (.701 vs. .764) stimuli actually improved somewhat on the delayed test.

In summary, the initial analysis confirmed the importance of category size on subsequent transfer, both immediately and after a delay of 1 wk. This analysis, however, does not address the issue of whether classification accuracy of new stimuli on the transfer test was influenced by old–new similarity. Figure 2 shows the mean classification accuracy of new transfer patterns as a function of their distances from old training stimuli (1.0–5.0 units/dot) for each category size and time of test. Also shown is the performance on the old and prototype stimuli. It should be noted that the prototype is, on the average, about 4.6 units/dot from old patterns. If the prototype were treated as simply a new pattern, then it should be classified no better than the most distant of the new patterns from an old stimulus (new patterns with old–new similarity of 5.0 units/dot).

An analysis based on old–new similarity (1.0–5.0 units/dot) for new stimuli was performed, with the variables of category size and time of test. As before, performance on the new patterns increased with an increase in category size, $F(2, 46) = 11.76$ (MS_e = 3.89), $p < .01$. The main effect of old–new similarity on classification of new patterns was highly significant, $F(4, 92) = 15.11$ ($MS_e = 1.04$), $p < .01$. The effectiveness of old–new similarity on transfer may be defined by the difference in classification accuracy for patterns that are 1.0 unit/dot versus those that are 5.0 units/dot from an old pattern. In general, the effectiveness of old–new similarity was diminished by both increasing category size and delay of the transfer test; for example, for the 5-instance category and an immediate test, the effectiveness of old–new similarity was 41%; for the 20-instance category on a delayed test, this value dropped to 7%. The interaction between old–new similarity and time of test was highly significant, $F(4, 92) = 8.64$ ($MS_e = 0.46$), $p < .01$, as was the Category Size × Old–New Similarity interaction, $F(8, 184) = 2.80$ ($MS_e = .77$), $p < .01$. In general, an increase of about 10% across the old–new similarity variable is necessary for statistical significance. Thus, the 6.7% difference due to old–new similarity for the 20-instance category on a delayed test was not significant, $F(4, 92) = 1.31$ ($MS_e = 0.59$), $p > .20$.

Figure 2 also reveals an interesting relationship between category size and old–new similarity on prototype classification. For the category defined by 5 exemplars, the prototype was classified more poorly than any type of new stimulus. In contrast, the prototype was classified better than any new pattern when the category was defined by 20 instances. For the category defined by 10 instances, the prototype was classified better than all new instances except for those that were very close to a particular old pattern. The importance of these results for exemplar-based versus prototype-based models of classification is discussed later.

[5] Since the opportunities of an error differed for the old (5), new (25), and prototype (2) stimuli, errors were manipulated by 20, 4, and 50, respectively, prior to the analysis. This had the effect of equating error opportunities for the different stimulus types, and produced a slightly more conservative test.

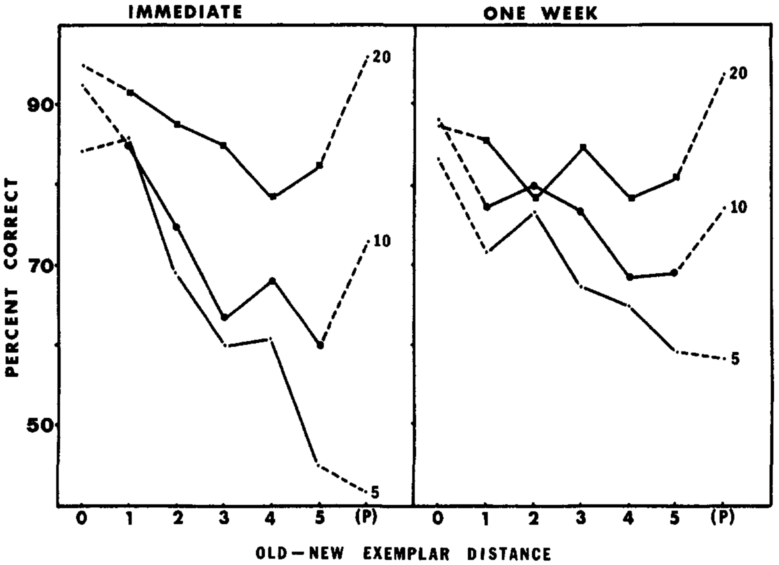

Figure 2. Mean classification accuracy of new stimuli on the transfer test, as a function of old-new similarity, category size, and time of test, Experiment 1.

Confidence ratings. Following the classification of each stimulus on the transfer test, the subjects were required to indicate how confident they were of their choice, with 1 = very uncertain, 2 = moderately confident, and 3 = very confident. Figure 3 shows the mean confidence ratings for correctly classified patterns, as a function of stimulus type, category size, and time of test. In addition, the mean confidence ratings for unrelated patterns that were erroneously classified into the learned categories are shown.

For the most part, the confidence ratings tended to mirror the classification performance. For example, the highest confidence ratings were obtained for the old stimuli, changes in confidence across category size were greatest for the prototype stimulus, and, overall, confidence ratings were reduced on the delayed test. Whereas the effect of old–new similarity was nicely ordered on an immediate test, it was less so on the delayed test, especially for the 20-instance category. For example, the ratings for the 20-instance patterns on the delayed test varied within a narrow range (2.25–2.60), with the highest ratings obtained for the prototype, and followed, in order, by the new-2 (new patterns with an old-new similarity of 2.0 units/dot), old, new-3, new-1, new-4, and new-5. This may be contrasted with the more orderly effect of old-new similarity on confidence for patterns in the 5- and 10-instance categories. Finally, the unrelated patterns were consistently accorded the lowest confidence ratings, independent of category size and time of test. Thus, whenever a new stimulus was properly assigned to its category, its associated confidence value far exceeded the confidence of an unrelated pattern that was assigned to the same category.

Classification of unrelated patterns. The likelihood that unrelated patterns were correctly assigned to the junk category was .461 on an immediate test and .442 on the delayed test. Unrelated patterns were erroneously classified into the 5-, 10-, and 20-instance categories with probabilities of .097, .167, and .275, respectively, on the immediate test; on the delayed test, these values were .117, .172, and .269, respectively. If the classification of unrelated patterns into the learned categories is viewed as a form of response bias (false alarm), then clearly a bias existed for those categories originally defined by more exemplars. However, corrections for bias, such as high-threshold models (Green & Swets, 1966), only reduce the magnitude

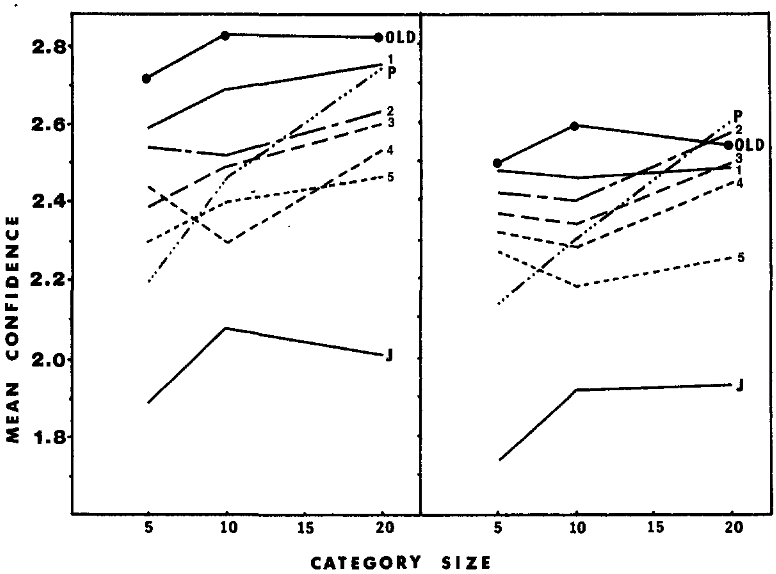

Figure 3. Mean confidence ratings for correctly classified patterns, as a function of stimulus type, old–new similarity, category size, and time of test, Experiment 1.

of the category size effect, and leave untouched the interaction of old–new similarity with category size. For example, the corrected values for new instances belonging to the 5-, 10-, and 20-instance categories were $p' = .620$, .666, and .772, respectively; the corresponding values for the prototypes were .440, .699, and .929, respectively.

There is some question, however, whether the classification tendencies of unrelated patterns should be used as an index of bias. First, the confidence ratings (Figure 3) indicate that the assignment of new and unrelated patterns that are sorted into the same category are clearly discriminated. Had subjects simply assigned more patterns into the larger categories on the transfer test, then it is unclear why the confidence ratings for correct assignments should markedly exceed the confidence ratings for erroneous assignments of unrelated patterns into the same category. Second, the correlation between hits and false alarms was computed for the 10- and 20-instance categories on both the immediate and delayed tests.[6] Contrary to what might be expected from a guessing interpretation, these correlations tended to be low and nonsignificant; for the 10-instance categories, these correlations were .33 (immediate) and .03 (delayed); for the 20-instance category, these values were .09 (immediate) and .42 (delayed), respectively. A complete summary of response tendencies for all stimulus types (old, new, prototype, unrelated) as a function of category size and time of test is shown in the left-hand panels of Table 2; the overall category response percentages are given on the bottom row. The implication of these response tendencies is discussed more fully in the general discussion.

Discussion

The results of Experiment 1 demonstrated that old–new similarity is an important factor influencing subsequent transfer, but that its influence is substantially diminished by increases in category size and a delay of the transfer test. Thus, the improved transfer performance obtained for the larger categories was associated with a diminished influence of old–new similarity. This result was obtained for both the accuracy data and the confidence ratings.

Nonetheless, one aspect of the procedure

[6] Correlations were confined to the 10- and 20-instance categories, because only these categories exhibited a reasonable range of false alarm values. Although the correlation between hits and false alarms was also low for the 5-instance category, the truncated false alarm range precluded an adequate test.

in Experiment 1 produced some bothersome concerns. A number of subjects reported at the conclusion of the study that the lengthy learning session (which sometimes lasted 1½–2 hr.) left them mentally exhausted by the time they participated in the immediate transfer test. If true, performance on the immediate transfer test may have underestimated the subject's knowledge about the acquired categories. The slight improvement on the delayed transfer test is consistent with this concern.

Experiment 2 was essentially a replication of Experiment 1, but a between-subject design was used to assess the effect of immediate and delayed transfer performance. In addition, the learning phase was slightly modified for those subjects in the immediate condition. On the first day, these subjects were brought to learning criterion. On the following day, these subjects were brought back to criterion and then administered the (immediate) transfer test. It was hoped that this procedural modification for the immediate subjects would provide an index of immediate transfer performance that was uncontaminated by fatigue.

Experiment 2

Method

Subjects. A total of 48 Arizona State University undergraduates served as subjects, half in an immediate transfer condition and half in a 1-wk.-delay condition.

Procedure. The procedure was identical to that used in Experiment 1, except that subjects in the immediate test condition received their learning and transfer tests on consecutive days. Specifically, all subjects in the immediate condition classified patterns to a criterion of two consecutive errorless trials on the first day, and were then dismissed with instructions to return the following day. On the following day, each subject again classified patterns to two consecutive errorless trials and then received the transfer test. With a few exceptions, most subjects received the minimum number of trials on the second day.

Design. A mixed design was used, with the variables of category size (5, 10, 20), stimulus type (prototype, old, new, random), and old–new similarity (1.0–5.0 units/dot) manipulated as within-subject variables, and time of the transfer test (immediate, 1-wk. delay) as a between-subject variable.

Results

Original learning. The mean number of trials to reach criterion was 14.33 for the immediate subjects and 13.04 for the delay subjects ($p > .05$). The trials of last error for categories represented by 5, 10, and 20 patterns were 9.54, 11.38, and 10.00, respectively, for the immediate subjects; for the delay subjects, the corresponding values were 8.96, 10.61, and 8.48, respectively. The apparent tendency for the intermediate-size category to be learned more slowly was significant, $F(2, 46) = 11.98$ ($MS_e = 4.14$), $p < .05$. Neither the effect of delay nor the Delay × Category Size interaction was significant.

Transfer performance. As was the case in Experiment 1, the main effects of category size, $F(2, 92) = 16.33$ ($MS_e = 568$), and stimulus type, $F(2, 92) = 12.87$ ($MS_e = 512$), were highly significant, as was the Category Size × Stimulus Type interaction, $F(4, 184) = 13.04$ ($MS_e = 363$), all $ps < .001$. Performance systematically improved with increases in category size (.729, .792, .829), with overall accuracy on the old patterns (.888) exceeding that of the new (.762) or prototype (.781) stimuli. The effect of increasing category size resulted in a 38% improvement for the prototype stimulus, 10% for new patterns, and 0% for the old stimuli. Contrary to the results of Experiment 1, performance on an immediate test (.826) exceeded that on the delayed test (.740), $F(1, 46) = 8.96$ ($MS_e = 1019$), $p < .01$. The size of the decrement was largest for items belonging to the 5-instance category (10%–16%) and smallest for the new and prototype stimuli belonging to the 20-instance category (0%–7%).

The mean classification accuracy of new transfer patterns at each level of distance from an old training stimulus (1.0–5.0 units/dot) is shown in Figure 4, as a function of category size and time of test. Also shown is the performance on the old and prototype stimuli.

An analysis of variance was performed on the new stimuli only, for the variables of old–new similarity, category size, and time of test. The main effects of category size and old–new similarity were highly significant, $F(2, 92) = 5.61$ ($MS_e = 2.71$), and $F(4, 184) = 38.52$ ($MS_e = 0.72$), respectively, both $ps < .01$. The effect of category size on subsequent transfer was minimal when the old–new similarity was high. However, as the old–new similarity decreased, the facil-

Table 2
Summary of Response Tendencies

		Category response															
		Experiment 1								Experiment 2							
		Immediate				Week delay				Immediate				Week delay			
Category membership	Stimulus type	5	10	20	None	5	10	20	None	5	10	20	None	5	10	20	None
5	Old	*.842*	.033	.042	.083	*.833*	.058	.058	.050	*.975*	.008	.017	.000	*.808*	.067	.100	.025
	New	*.642*	.065	.108	.185	*.680*	.085	.097	.138	*.782*	.038	.075	.105	*.635*	.130	.167	.068
	Prototype	*.417*	.083	.271	.229	*.583*	.146	.062	.208	*.625*	.104	.167	.104	*.521*	.208	.146	.125
10	Old	.008	*.925*	.050	.017	.017	*.883*	.075	.025	.008	*.900*	.058	.033	.042	*.850*	.108	.000
	New	.030	*.703*	.100	.167	.040	*.743*	.112	.104	.028	*.787*	.107	.078	.050	*.757*	.138	.055
	Prototype	.000	*.729*	.167	.104	.000	*.771*	.104	.125	.021	*.917*	.000	.062	.062	*.729*	.188	.021
20	Old	.000	.008	*.950*	.042	.008	.042	*.875*	.075	.000	.042	*.942*	.017	.025	.083	*.850*	.042
	New	.008	.037	*.850*	.105	.023	.037	*.817*	.123	.018	.057	*.842*	.083	.053	.083	*.770*	.093
	Prototype	.000	.042	*.958*	.000	.000	.042	*.938*	.021	.000	.000	*.938*	.062	.000	.042	*.958*	.000
None	Unrelated	.097	.167	.275	*.461*	.117	.172	.269	*.442*	.083	.192	.292	*.433*	.142	.192	.328	*.339*
	Total	.212	.263	.348	.178	.232	.280	.333	.159	.254	.286	.336	.125	.235	.307	.357	.100

Note. Italicized values signify correct category responses.

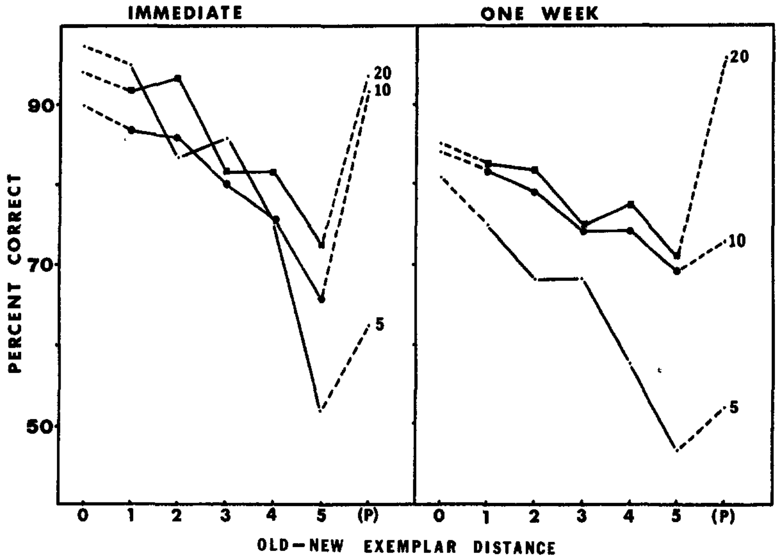

Figure 4. Mean classification accuracy on new stimuli on the transfer test, as a function of old–new similarity, category size, and time of test, Experiment 2.

itative effect of increased category size became more pronounced. For example, the mean classification accuracy of new patterns that were minimal distortions of an old pattern (new–1) was .850 for the 5-instance category and .871 for the 20-instance category. When new patterns were substantial distortions of an old stimulus, the effects of category size were substantial. For example, for new–4 patterns, the classification accuracy was .662 and .796 for the 5- and 20-instance categories, respectively; for new–5 patterns, the corresponding values were .492 and .716. As expected, the Category Size × Old–New Similarity interaction was significant, $F(8, 368) = 4.18$ ($MS_e = .67$), $p < .01$.

Overall, classification accuracy for the new patterns decreased by 8.4% across the week delay, $F(1, 46) = 8.41$ ($MS_e = 3.77$), $p < .01$. The transfer decrement was greater for patterns belonging to the 5-instance category (15.0%) than either the 10- or 20-instance categories (3.0% and 7.9%, respectively). The transfer decrement across the week delay also tended to be reduced by increases in old–new dissimilarity; for the new–1, new–2, new–3, new–4, and new–5 patterns, the magnitude of this decrement was 10.7%, 10.3%, 11.6%, 8.6%, and 1.9%, respectively. However, the interaction between old–new similarity and time of the transfer test fell just short of significance, $F(4, 184) = 2.36$ ($MS_e = 0.72$), $.05 < p < .10$.

Finally, performance on the prototype mirrored the results obtained in Experiment 1. Specifically, (a) classification accuracy for the prototype was markedly enhanced by increases in category size, the magnitude of this facilitation being 31% on the immediate test and 41% on the delayed test; and (b) the prototype was classified more poorly than most new patterns when the category was defined by only 5 patterns; when the category was defined by 20 exemplars, the prototype was classified more accurately than any new pattern.

Confidence ratings. Figure 5 shows the mean confidence ratings for correctly classified patterns, as a function of stimulus type, category size, and time of test. As was the case in Experiment 1, the mean confidence ratings were nicely ordered according to old–new similarity for patterns belonging to the 5- and 10-instance categories on an immediate test, with the old stimuli receiving the highest ratings and the prototype receiving the lowest ratings for the 5-instance category. This systematic ordering becomes

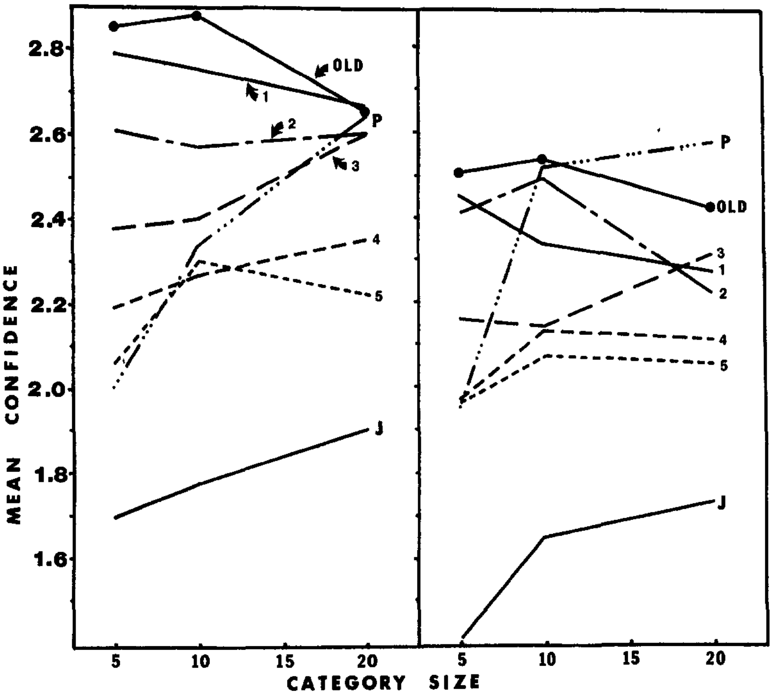

Figure 5. Mean confidence ratings for correctly classified patterns, as a function of stimulus type, old–new similarity, category size, and time of test, Experiment 2.

somewhat muddled for the 20-instance category on the immediate test, and for both the 10- and 20-instance category on the delayed test. The mean confidence ratings for the prototype belonging to the 20-instance category exceeded the mean confidence for the old patterns and all new patterns on the delayed test. Unrelated patterns, when (erroneously) classified into the learned categories, were clearly the recipient of the lowest confidence ratings. Unlike the results of Experiment 1, the mean confidence ratings were reduced for all stimulus types on the delayed test, with the old patterns and new-1 and new-2 patterns showing the greatest decrement.

Classification of unrelated patterns. The likelihood that unrelated patterns were correctly assigned to the junk category was .433 on the immediate test and .339 on the delayed test. Unrelated patterns were erroneously classified into the 5-, 10-, and 20-instance categories with probabilities of .083, .192, and .292, respectively, on the immediate test; on the delayed test, these values were .142, .192, and .328, respectively.

The correlation between false alarms and hits for the 10-instance category was −.13 (immediate) and +.16 (delay); for the 20-instance category, these correlations were +.12 (immediate) and −.02 (delay). None of these correlations approached significance. A summary of response tendencies for all stimulus types, as a function of category size and time of test, is shown in the right-hand panels of Table 2.

General Discussion

Two major results emerged from the present study: (a) Increases in the number of exemplars that represented a category in the learning phase resulted in a marked improvement in the classification of novel patterns belonging to that category on a later transfer test; and (b) the likelihood that a novel pattern was categorized by its similarity to an old pattern was increasingly attenuated by increases in category size and delay of the transfer test. The former result has been obtained repeatedly (e.g., Homa, 1978) and is consistent with the view that the mental representation of a category evolves with increasing exemplar experience. The latter result is germane to the issue of what determines generalization in a categorization task. As noted previously, exemplar-based models of generalization that posit the retrieval of a single most similar exemplar (Hintzman & Ludlam, 1980) or a fixed number of exemplars at the time of classification must predict that classification is an additive function of old–new similarity and category size. Since these variables strongly interacted in both experiments, exemplar models of this type are inadequate to account for the present results. Prototype models that assume that only the central tendency is stored can also be rejected as explanations for the present results, at least when categories are defined by relatively few exemplars, since objective old–new similarity was a significant determinant of classification. Only the mixed prototype model and the complete set exemplar model seem potentially compatible with the present results, since both models can predict an interaction between category size and old–new similarity.

The suitability of these two models was further assessed by noting the nature of the quantitative fit of Equation 1 to the observed classification data. According to Medin and Schaffer (1978), x varies from .00 to 1.00 and represents the similarity contribution to classification of an old pattern to a transfer pattern. At issue is whether the similarity contribution to classification of specific old–new similarity is independent of category size and time of test (complete set exemplar model) or whether the contribution of specific old–new similarity is diminished by increases in category size and length of time between acquisition and transfer (mixed prototype model). In one case, the x values were derived separately for each category size at each level of objective old–new similarity such that a perfect fit of the classification data resulted. In another case, a best fit of the classification data was computed for a set of x values common to category size. The concern was whether the goodness of fit seemed adequate to account for the present results, and whether the magnitude of the observed category size effect was accounted for by Equation 1.

For the first case, three different sets of derived x values were determined, each under slightly different assumptions: For Condition 1, assume $x_o = 1.00$, $s_w = .01$, and s_b is estimated separately for each category size and time of test; for Condition 2, assume $x_5 = .05$, $s_w = .01$, and s_b is estimated separately for each category size and time of test; and for Condition 3, assume $s_w = .01$, and hold s_b constant across category size and time of test, with x unbounded. Here, x_o represents the similiarity contribution to classification of an old pattern presented at transfer to its stored representation, and $x_1, \ldots, x_5$ represent the estimated contributions to classification for new patterns that objectively differed by 1.00 unit/dot, . . . , 5.00 units/dot, respectively, from a stored old pattern. In Condition 1, the x values are derived under the assumption that an old pattern on a transfer test has a similarity value of 1.00 with its stored representation. The value of s_w is set equal to .01 for all category sizes and times of test for two reasons. First, $s_w = .01$ is about the maximum value that is permissible for all category sizes and times of test, given that x must vary between .00 and 1.00. Second, the objective distance of new patterns to all training patterns (except for the old pattern that was at a manipulated distance to a new pattern) was roughly equal for each category size. By estimating s_b separately for each category size and time of test, it was possible to derive values of x that perfectly predicted the classification data. In Condition 2, the x values are derived under the assumption that the most distant of the manipulated levels of old–new objective similarity (old–new = 5.00 units/dot) results in a low value of x that is constant across category size and time of test. The assumed value of x_5 still exceeds that for s_w (.05 vs. .01). Thus, the contribution to classification of the most extreme level of manipulated old–new similarity is still greater than that of two arbitrarily selected patterns within the same category. This assumption is consistent with the fact that the objective interpattern distance for two arbitrary patterns within the same category is greater (7.50 units/dot) than for patterns at the most extreme level of old–new similarity (see Table 1). In Condition 3, a single value of s_w and s_b is selected for all category sizes and times of test, and x is allowed to assume whatever value is needed to fit the classification data (including negative values).

The estimated values for x under these three conditions are shown in Table 3, as a function of category size, time of test, and objective old–new similarity. The derived x values shown in Table 3 represent only a sample of the conditions that have been explored. The classification data for Experiment 1 are used for Conditions 1 and 2; the classification data for Experiment 2 are used for Condition 3.

For all three conditions, the derived x values interact with category size and tend to become uniform across objective old–new similarity on the delayed test. In all cases, the greatest values of x are obtained for the 5-instance category, and the smallest values are generally associated with the 20-instance category.[7] It should be noted that Condition 1 effectively pegs the x values at the most similar of the manipulated levels of similarity (x_o = 1.00), whereas Condition 2 pegs the x values at the most dissimilar level of old–new similarity (x_5 = .05). There exists a family of curves relating x to old–new similarity for the variables of category size and time of test that are intermediate to Conditions 1 and 2. None of these functions would alter the fact that the derived values of x are an

[7] The single exception occurs in Condition 1 on the delayed test, where the x values for the 20-instance category are quite large. This outcome was produced by the unusually large value of s_b (.0113) that was needed to fit the classification data. In fact, s_b (the similarity between patterns in different categories) has a slightly larger value than s_w (the similarity between patterns in the same category).

Table 3

Estimated Values of x *as a Function of Old–New Similarity, Category Size (5, 10, 20), and Time of Test, Under Three Different Sets of Assumptions (1, 2, 3)*

	Values of x								
	1 (x_o = 1.00)			2 (x_5 = .05)			3 (x unbounded)		
Objective similarity	5	10	20	5	10	20	5	10	20
Immediate									
Old–new									
1.0	1.18	.40	.51	.64	.43	.37	1.10	.24	.14
2.0	.40	.17	.25	.21	.19	.17	.26	.21	.23
3.0	.24	.06	.17	.12	.07	.10	.32	.11	−.06
4.0	.28	.10	.04	.14	.11	−.01	.14	.06	−.06
5.0	.12	.04	.11	.05	.05	.05	.02	.01	−.11
Old–Prototype	.03	.02	.07	.02	.02	.06	.02	.06	.02
Delayed									
Old–New									
1.0	.49	.41	.83	.12	.12	.15	.14	.13	−.05
2.0	.65	.49	.42	.17	.16	.02	.09	.10	−.06
3.0	.40	.39	.77	.09	.12	.13	.09	.05	−.10
4.0	.35	.22	.42	.08	.04	.02	.04	.05	−.09
5.0	.26	.24	.52	.05	.05	.05	.01	.02	−.12
Old–Prototype	.06	.05	.13	.02	.02	.04	.01	.01	.03
s_w	.0100	.0100	.0100	.0100	.0100	.0100	.0100	.0100	.0100
s_b: immed.	.0066	.0035	.0042	.0037	.0037	.0034	.0020	.0020	.0020
s_b: delay	.0070	.0058	.0113	.0021	.0025	.0038	.0020	.0020	.0020

Note. s_w = average within-category similarity; s_b = average between-category similarity.

interactive function of category size and time of test. In Condition 3, the estimated x values are shown when s_b has been set equal to .002. However, no value of s_b would alter the basic patterning of results; the values of x again interact with category size and, on the delayed test, become relatively uniform across old–new similarity. Taken at face value, these results indicate that the contribution of specific old–new similarity is diminished by the variables of category size and time of test.

Still, it may be argued that the parameter space in the present study is not sufficiently sharp to warrant conclusions regarding the derived x values. To assess this possibility, a best fit of the classification data was computed with x values common to category size.[8] Table 4 shows the results of a best fit to the data of Experiment 1, immediate test. Generally, the complete set exemplar model provides a reasonable fit to the observed data, predicting a strong effect of objective old–new similarity for the 5-instance category and a diminished effect of old–new similarity for the 20-instance category. The major shortcomings of the best fit are shown in Table 5, which summarizes the predicted and observed category size effects for this set of data. Here, the category size effect is computed separately for old, new, and prototype patterns and reflects the difference in classification accuracy for the 5-instance and 20-instance categories. As indicated in Table 5, the predicted benefits of category size for old patterns (.067) are somewhat less than observed (.108). More bothersome, however, is the fact that the predicted category size effect is substantially smaller than observed for new and, especially, for prototype patterns. In fact, when best fits are computed separately for each of the four data sets (Experiment 1, immediate and delayed test; Experiment 2, immediate and delayed test), a similar outcome is obtained: In each case, the predicted magnitude of the category size effect is smaller than observed for new and prototype patterns, whereas the predicted category size effect is usually overestimated for the old patterns (Experiment 1, immediate test, is the lone exception).[9]

Table 4
Best Fit of Classification Performance for Experiment 1, Immediate Test

	Category size					
	5		10		20	
Estimated value of x	Predicted	Observed	Predicted	Observed	Predicted	Observed
x_o = 1.00	.874	.842	.897	.925	.941	.950
x_1 = .72	.835	.858	.866	.850	.924	.917
x_2 = .30	.694	.692	.757	.750	.867	.875
x_3 = .17	.583	.600	.675	.633	.828	.850
x_4 = .16	.571	.608	.667	.683	.824	.783
x_5 = .10	.483	.450	.603	.600	.795	.825
x_p = .03	.500	.417	.706	.729	.889	.958

[8] Best fits were computed separately for each of the four data sets and involved the estimation of 8 parameters (s_w, s_b, x_o, . . . , x_5 and x_p) to 21 data points (3 Category Sizes × 7 Transfer Items: old, prototype, and 5 levels of new). The only restrictions were that the x values be bounded between .00 and 1.00, and that the estimated x values decrease with increases in objective old–new similarity.

[9] Ideally, the parameter values resulting from a best fit to data will be theoretically meaningful. If one were to accept Equation 1 as appropriate to the present study, then the resulting parameter values describe a complex, if not peculiar, set of influences that determined classification performance. For both experiments, the value of s_b increased on the delayed test, suggesting that between-category discriminability was worsened by a week's delay. However, the best-fitting values of x_1–x_5 interacted with time of test, with the result that the x values for moderate and extreme levels of objective old–new similarity were considerably larger on the delayed test. For example, in Experiment 1, immediate test, the values of x_1–x_5 were .72, .30, .17, .16, and .10, respectively. On the delayed test, the corresponding values were .57, .56, .50, .34, and .33. A similar result, in which x values again interacted with time of test, was obtained in Experiment 2. On the one hand, we would have to

Table 5
Predicted and Observed Category Size Effects

Stimulus type	Predicted	Observed	Error
Old	.067	.108	−.041
New	.223	.316	−.093
Prototype	.389	.541	−.152

The consistency between the two quantitative approaches can now be appreciated. For the data to be perfectly predicted (Table 3), it is necessary to have x values interact with category size and time of test; when a common set of x values are selected that produce a best fit to the classification data (Table 4), the magnitude of the predicted category size effect is, for new and prototype patterns, consistently underestimated. These two outcomes can only be realized by assuming that the contribution to classification of specific old–new similarity is increasingly diminished by increases in category size and delay of test.

The overestimation of the category size effect for old patterns by Equation 1 deserves comment. In abstraction research, the transfer performance of old patterns is usually invariant across category size; that is, the category size effect is typically of small magnitude or absent totally (e.g., Homa et al., 1973; Homa & Hibbs, 1978). This outcome is hardly surprising, since transfer is assessed only following errorless classification in the learning phase. In the present study, which also used an errorless learning criterion, the magnitude of the observed category size effect for old patterns was .108 and .042 in Experiment 1, immediate and delayed test, respectively; in Experiment 2, these values were −.033 and +.042. Thus, with the exception of performance in Experiment 1, immediate test, the observed magnitude of the category size effect was quite small, averaging less than a 2% facilitation. In contrast, the predicted magnitude of the category size effect by Equation 1 was sizable, averaging 8.5%, with a range of 6.7% to 10.9% for the four data sets. The predicted category size effect arises because the denominator of Equation 1 cumulates dissimilarities from the contrasting categories; the larger the size of the contrasting categories, the poorer the predicted performance on old patterns from the smallest category will be. In fact, for the best fitting parameters in the present study, an upper bound of 87% is predicted for old patterns belonging to the smallest category. This upper limit was substantially exceeded in Experiment 2, immediate test (observed = .975; predicted = .874), resulting in the category size effect being overestimated by 10%.

Since the complete set exemplar model has no assumptions to distinguish the classification of old and new patterns, a category size effect is predicted for both stimulus types. As a consequence, it seems inevitable that the category size effect for old patterns will be overestimated by Equation 1 whenever an errorless learning criterion is adopted, especially when the categories are also defined by numerous exemplars. It is not clear that Equation 1 could be modified to rectify this problem. For example, the magnitude of error for old patterns could be reduced by simply assuming that a near-equal number of patterns is stored for each category, regardless of how many patterns actually defined each category. Generally, this would reduce the predicted magnitude of the category size effect for old patterns to be more in line with the obtained values. However, modifications of this type would only further increase the error for new and prototype patterns; that is, the predicted category size effect for these patterns would also be reduced, thereby underestimating the category size effect by an even greater amount. Alternatively, Equation 1 could be modified to include a familiarity or strength parameter

explain why the x values associated with minimal old–new differences declined on the delayed test, whereas the x values for the largest old–new differences increased. On the other hand, we would have to explain why the generally elevated x values on the delayed test occurred, given that the *degree* of specific old–new similarity was considerably less important on the delayed test; that is, the x values are near asymptote across objective old–new similarity. Given the occurrence of these events within the context of a general deterioration of between-category discriminability, it is as if the entire categorical space were decaying to an unlearned state, and all new patterns were located equidistant from their designated old patterns.

for old patterns that would be at asymptote immediately and decay thereafter. By setting the strength parameter equal for all category sizes at the conclusion of learning, predicted differences across category size for old patterns would be reduced. However, assumptions of this type would violate the spirit of current versions of exemplar theory that assumes that the same mechanism mediates classification of old and new patterns (Medin & Schaffer, 1978). Regardless, even this modification would leave untouched the underestimation of the category size effect for new and prototype patterns.

In sum, the results of the present study are more compatible with a mixed prototype model. There are two major shortcomings of an exemplar-based model of generalization such as is expressed in Equation 1; the model *underestimates* the observed category size effect for the category prototype and, to a lesser extent, new patterns, and it generally *overestimates* the observed category size effect for old patterns. The patterning of results shown in Table 3 suggests that transfer to specific, stored exemplars may be most likely when the category is represented by a small number of exemplars, an immediate test is administered, and the similarity between a novel stimulus and a stored exemplar is high. However, exemplar-based generalization is increasingly unlikely once exemplar experience becomes substantial and a delayed test is used. Previous support for exemplar-based generalization was obtained under conditions of minimal exemplar experience and an immediate transfer test (Medin & Schaffer, 1978). A reasonable hypothesis is that a concept, in its early stage of development, is represented primarily by a small number of exemplars. With continued learning, however, the concept becomes increasingly represented by the central tendency (abstracted prototype) and the breadth of that concept. Consistent with this view is the fact that variable manipulations that enhanced the classification of the prototype were also associated with a diminished importance of old–new similarity on classification. One obvious implication of the present results is that the importance of exemplar similarity on generalization may asymptote to zero as category experience continues to increase.

The theoretical importance of increased exemplar experience on transfer for classification models cannot be stressed enough. The human organism encounters an essentially limitless array of examples that belong to biological, invented, and esthetic categories. It has even been proposed that the beginning scientist is able to appreciate the abstract concepts and paradigmatic rules of a profession only after repeated laboratory demonstrations (Kuhn, 1970), that is, by numerous, concrete examples. The suggestion here is that the nature of categorical information that guides generalization behavior (specific exemplars or an abstracted prototype) interacts with the amount of exemplar experience relevant to a category.

A number of secondary results are of interest. First, the category prototype need not be classified any better than other new instances, especially when category experience is minimal and only high-level distortions represent a category. In both Experiments 1 and 2, the objective prototype was classified more poorly than most new stimuli when the category size for that category was small (e.g., 5 exemplars). The advantaged position of the prototype on the transfer test was realized only when the category was defined by numerous exemplars. Second, confidence ratings provided a good mirror of classification accuracy. For example, confidence ratings declined across the delay, especially for old stimuli and new patterns that shared a high degree of similarity to the old stimuli. The mean confidence ratings tended to be more stable for the prototype and new patterns that shared a minimal similarity to the old patterns. Third, the overall performance on the transfer test was influenced by the type of experimental design used (within- vs. between-subject). Specifically, less forgetting was obtained when a within-subject design was used (Experiment 1), a result inconsistent with earlier findings in category abstraction (Strange et al., 1970), which had found no retention differences. The source of this discrepancy is unclear, although the present study would seem to be a more difficult task (35 high-level distortions had to be classified in the learning phase vs. 12 distortions in the experiment by Strange et al.).

Two other issues warrant discussion. First, exemplar-based generalization was assessed in the present study by the manipulation of *objective* similarity between old and new patterns. Is it possible that the *subjective* similarity between old and new exemplars determined transfer performance, even for the largest categories in which objective similarity played a negligible role? Second, why did increased exemplar experience in the learning phase result in a greater tendency to classify unrelated patterns into the category on the transfer test? With regard to the first issue, the present study cannot directly address the nature of the internal representation of the category members. Nonetheless, the objective similarity between old and new stimuli was an important predictor of transfer performance, but primarily when the category size was small. If the subject stored intact patterns but the encoded representation failed to preserve the objective old–new similarity, then the objective manipulation of old–new similarity should have been ineffective for all category sizes at each delay. Since this was not the case, exemplar-based models of classification would have to explain why an effective manipulation of objective old–new similarity interacted with category experience and time of test.

The question of why overgeneralization appears to occur for those categories that show the greatest levels of abstraction is unclear. A simple explanation is that the subject simply increases his willingness to classify patterns into that category originally defined by the most patterns in learning (a bias explanation). An alternative explanation is that increased levels of abstraction produce, as an initial by-product, a categorical breadth that is overly extended. An explanation for overgeneralization that is based solely on bias seems unlikely for a number of reasons. First, a bias explanation would predict that a subject's hit and false alarm rate should be positively correlated. However, the correlation between hits (correct assignments of new patterns) and false alarms (erroneous classifications of unrelated patterns into the learned categories) was low and nonsignificant in the present study. In effect, a subject who gave evidence of strong abstraction tendencies by sorting nearly all the new patterns into the appropriate category was as likely to sort few unrelated patterns into that category as many. Second, subjects clearly discriminated between the new and unrelated patterns that were classified together into the same category, as judged from the confidence ratings. If transfer performance were corrected by including only those new patterns that were correctly classified and that exceeded the confidence value for unrelated patterns sorted into that category, then the magnitude of the category size effect would be virtually unchanged. What is most puzzling is the fact that a subject who clearly discriminates new from unrelated patterns will nonetheless sort both stimulus types into the same category. A partial answer is that some subjects view the unrelated patterns as extreme distortions of the category, that is, make an overgeneralization based on the information contained in the pattern rather than a criterion shift.

Two results from previous research also argue against a bias interpretation: (a) It is not unusual to obtain a category size effect when differential bias is absent (e.g., Homa, 1978, Experiment 2; Homa & Hibbs, 1978; Omohundro, 1981); and (b) the magnitude of the category size effect on subsequent transfer is unaffected by either explicit (Homa & Field, Note 2) or implicit (Omohundro & Homa, 1981) manipulations of a supposed criterion. In sum, the evidence suggests that manipulations that enhance the degree of abstraction for a category may also result in a representation that is overly extended. It may well be the case that only manipulations that emphasize category discriminability (e.g., training on numerous rather than few categories, Homa & Chambliss, 1975) can effectively counteract tendencies toward overgeneralization.

Reference Notes

1. Homa, D., Omohundro, J., & Courter, S. *Perception of abstracted form.* Paper presented to the Rocky Mountain Psychological Association, Albuquerque, N.M., May 1977.

2. Homa, D., & Field, D. *Breadth and bias in category abstraction.* Paper presented to the Psychonomic Society, St. Louis, Mo., November 1980.

References

Brooks, L. Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization.* Hillsdale, N.J.: Erlbaum, 1978.

Fisher, S. C. The process of generalizing abstraction; and its product, the general concept. *Psychological Monographs,* 1916, *21*(2, Whole No. 90).

Green, D. M., & Swets, J. A. *Signal detection theory and psychophysics.* New York: Wiley, 1966.

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

Hintzman, D. L., & Ludlam, G. Differential forgetting of prototypes and old instances: Simulation by an exemplar-based classification model. *Memory & Cognition,* 1980, *8*, 378-382.

Homa, D. Abstraction of ill-defined form. *Journal of Experimental Psychology: Human Learning and Memory,* 1978, *4*, 407-416.

Homa, D., & Chambliss, D. The relative contributions of common and distinctive information on the abstraction from ill-defined categories. *Journal of Experimental Psychology: Human Learning and Memory,* 1975, *1*, 351-359.

Homa, D., Cross, J., Cornell, D., Goldman, D., & Schwartz, S. Prototype abstraction and classification of new instances as a function of number of instances defining the prototype. *Journal of Experimental Psychology,* 1973, *101*, 116-122.

Homa, D., & Hibbs, B. Prototype abstraction and the rejection of extraneous patterns. *Bulletin of the Psychonomic Society,* 1978, *11*, 1-4.

Homa, D., Rhoads, D., & Chambliss, D. The evolution of conceptual structure. *Journal of Experimental Psychology: Human Learning and Memory,* 1979, *5*, 11-23.

Homa, D., & Vosburgh, R. Category breadth and the abstraction of prototypical information. *Journal of Experimental Psychology: Human Learning and Memory,* 1976, *2*, 322-330.

Hull, C. L. Quantitative aspects of the evolution of concepts. *Psychological Monographs,* 1920, *28*(1, Whole No. 123).

Kuhn, T. S. *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press, 1970.

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

Neisser, U. *Cognitive psychology.* New York: Appleton-Century-Crofts, 1967.

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

Omohundro, J., & Homa, D. Search for abstracted information. *American Journal of Psychology,* 1981, *94*, 267-290.

Posner, M. I., Goldsmith, R., & Welton, K. E., Jr. Perceived distance and the classification of distorted patterns. *Journal of Experimental Psychology,* 1967, *73*, 28-38.

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

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

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

Strange, W., Kenney, T., Kessel, F. S., & Jenkins, J. J. Abstraction over time of prototypes from distortions of random dot patterns: A replication. *Journal of Experimental Psychology,* 1970, *83*, 508-510.

Received December 30, 1980
Revision received June 29, 1981 ■

Notice of Journal Title Change

By action of APA's Publications and Communications Board, the title of the

Journal of Experimental Psychology: Human Learning and Memory

will be changed to

Journal of Experimental Psychology: Learning, Memory, and Cognition

as of the January issue of the 1982 volume (Vol. 8, No. 1). The volume numbering will continue from the last volume published under the former title, and the journal will continue to be published bimonthly, in one volume per year.

This title change reflects the journal's expanded coverage in the area of human cognition. For further information on content and submission of manuscripts, authors should refer to the editorial in the July 1980 (Vol. 6, No. 4, pp. 439-440) issue of *JEP: Human Learning and Memory* and to the Instructions to Authors, which are published in each issue of the journal.