Journal of Experimental Psychology: Animal Behavior Processes

VOL. 2, NO. 4 OCTOBER 1976

Natural Concepts in Pigeons

R. J. Herrnstein, Donald H. Loveland, and Cynthia Cable
Harvard University

Pigeons learned to discriminate pictures of trees, bodies of water, or a particular person in three separate experiments. Pictures being seen for the first time were discriminated almost as well as pictures seen in training. The pigeons in each experiment showed similar patterns of errors and correct discrimination.

The list of stimuli used in typical discrimination-learning experiments may be long, but it is characterized by a definiteness that sets it apart from the list of stimuli animals confront in nature. In nature, open-ended variability is the rule rather than reproducibility or definiteness. Squirrels may learn where the acorns are, but the acorns will vary in size, shape, and color, and so will the oak trees. Mice may learn to be wary of houses and yards with cats in them, even though cats look different. The variation within naturally occurring stimulus classes typically defies our capacity for physical definition, unlike the 1,000-Hz tones or 465-mμ lights of a psychological experiment. It is doubtful whether anyone can yet instruct a machine to identify acorns or cats let alone the stimulus classes that form the basis of human language, such as *chair, house,* or *mama.*

There seem to be two sorts of reasons why experimentation fails to match nature better—the practical and the theoretical. Practically speaking, it is far easier to present a rat with repeated bursts of a 1,000-Hz tone than to reproduce the limitlessly variable sounds it responds to in nature—the peeps of other rats, for example. Similarly, it is easier to present pigeons with a red circle of specified diameter than to take hundreds or thousands of pictures of the roughly spherical grains it eats.

The practical bias toward reproducibility is bolstered by theory. Natural stimulus classes are held to be merely imperfect analogues of the artificial stimulus classes of the laboratory. On the origin of stimulus classes, theorists as different in other respects as Hull (1943), Skinner (1953), and Bruner (Bruner, Goodnow, & Austin, 1956) fall mainly in the classical empirical tradition. Stimulus classes supposedly arise in the process of induction. For reinforcement theorists like Hull and Skinner, reinforcement produces the induction; for cogni-

The research reported here was supported by a grant from the National Aeronautics and Space Administration to Harvard University. Preparation of the manuscript and data analysis were partially supported by Grant MH-15494 from the National Institute of Mental Health to Harvard University. In addition, the authors would like to express special thanks to John Cerella for the development of the computer program that controlled the experimental procedures, to Arlene A. Pippin and Virginia Koster for valuable help in the analysis of data, and to David Green for comments on the relations between our analysis and signal-detection theory.

Cynthia Cable is now at 720 North Avenue 65, Los Angeles, California 90042.

Requests for reprints should be sent to R. J. Herrnstein, Department of Psychology and Social Relations, William James Hall, Harvard University, Cambridge, Massachusetts 02138.

tivists like Bruner, the induction is taken as given. Either way, the creature sees, for example, a series of acorns and nonacorns which supposedly yields the defining parameters of acorns in general. Some set of values on a collection of visual variables marks off the region of acorns.

It is hard, and expensive, to study discrimination learning with the open-ended series of instances found in nature. It has been easier to assume that the *process* of discrimination is separable from the things being discriminated, even though there is reason to examine the possibility that, at least sometimes, the two are inseparably linked. But that possibility seems to require bringing natural discrimination into the laboratory. An approximation to such a natural discrimination procedure was reported by Herrnstein and Loveland (1964) and has since been replicated by others (e.g. Malott and Siddall, 1972). Pigeons were trained to peck a key only when they saw pictures containing people. They learned the discrimination rapidly and well, responding differentially to pictures seen for the first time. The essential feature of a natural discrimination—which is the ability to cope with natural ranges of variation—was at least approached. The present experiments extend the earlier findings by using other classes of stimuli.

Method

Subjects

Eleven male pigeons at about 80% ad-lib weight were used in three experiments. Three pigeons were homers (51H, 56H, and 63H); 8 were white Carneaux (7C, 8C, 24C, 44C, 45C, 91C, and 244C). Each pigeon worked in only one of the three present experiments. They all had previous experience in similar experiments with stimuli other than the ones reported here. Since there was no apparent differential effect of the earlier training on the present data, nothing more will be said about them.

Apparatus

A standard pigeon chamber was adapted for the projection of 35-mm slides by a Kodak Carousel projector. In addition to the standard pigeon key and feeder, the front wall of the chamber contained a rectangular screen 1¾ in. (.044 m) high × 2½ in. (.064 m) wide through which the slides were projected. The screen was in the center of the front wall, at the same height as the key but about 2 in. (.051 m) to its right. It took a force of about 15 g to operate the key. The feeder presented food for the 2.5 sec at a time, during which the key was darkened. A masking noise sounded continuously, and every peck of the key caused an audible click. An electromechanical circuit did the programming and recording for the early sessions, but then control was shifted to a PDP 9/T computer (Digital Equipment Corporation). All the data reported here came from the computer-run phase.

Procedure

During an experimental session, 80 slides were successively projected on the screen. The average duration of each presentation was 30 sec, varying irregularly from 10 to 90 sec. A 10-sec interval separated successive slides, during which the chamber was blacked out completely. The key was illuminated with a white light that came on 5 sec after the onset of a picture and stayed on (except during feeder operations) until the picture went off. Pecking counted only when the key was lit.

Each set of 80 pictures comprised 40 ± 5 examples of some object (e.g., trees, see below) and 40 ± 5 nonexamples. These are called positive and negative stimuli, respectively. The order of positive and negative stimuli was random and rescrambled for most sessions. That is to say, the pigeons had no opportunity to learn a given sequence of positive and negative stimuli, since most sessions presented new orders of pictures, even when the pictures themselves had been presented before.

In the presence of positive stimuli, pecking was reinforced on a variable-interval schedule with an average of 30 sec and a range from 3 to 90 sec. The schedule of reinforcement was independent of the schedule of stimulus durations even though their parameters were similar. The schedules permitted anywhere from zero to three reinforcements for a single positive stimulus. In the presence of negative stimuli, pecking earned no food. Moreover, a negative stimulus could not be terminated within 10 sec of a peck. This penalty for negative-stimulus pecking was superimposed on the basic schedule of stimulus duration.

The three experiments were identical except for the stimulus classes involved. In Experiment T, the positive class comprised pictures containing one or more trees or any part of a tree. In Experiment W, it was pictures containing water. In Experiment P, it was pictures containinng all or part of a particular person, a women in her late twenties who consented to having a photographer follow her from time to time over a period of about a year, taking posed and unposed pictures.

In all three experiments, the photographer (the third author) used a Nikon 35-mm single-lens reflex camera with a variety of lenses to take pictures during all of New England's four seasons. Experiment T (trees) employed 1,840 different pictures, of which half were positive and half negative. The trees were of virtually every description and variety found locally and were photographed from near or far, unobstructed and partially obstructed. For positive stimuli, we attempted to capture the full range of scenes considered to contain a tree or any part thereof, but not necessarily photographing the scene as if the tree were the center of attention. The sole restriction was that none of the 1,840 pictures contained shrubs or bushes. We chose to solve the problem of ambiguity here by avoiding it. The negative stimuli were comparable to the positive in every respect except for the presence or absence of trees, as far as we could tell.

Experiment W (water) used 1,760 different pictures, of which half were positive and half negative. Water meant anything from an aerial view of the Atlantic Ocean to obscure, small puddles. No pictures contained visible drops of water, as would be seen, for example, in rain. The negative stimuli often mimicked the more obvious visual properties of water—its shininess, smoothness, blueness, and so on. A number of pictures contained snow or ice. If these contained no liquid water, they were designated negative.

Experiment P (specific person) used 1,600 different pictures, of which half were positive and half negative. As before, every attempt was made to capture the person in the full range of settings—indoors or outdoors, near or far, front or rear, clear or obscure, alone or with other people.

All pictures were in color but of varying brightness and contrast. Many pictures served in more than one experiment, often switching status as negative or positive but not always. The pictures were *not* composed to draw attention to the relevant features. That is to say, many positive stimuli had the critical element off center, small, or distant, and therefore were easily overlooked by casual human observers. The decision about the contents of each picture was usually made by the photographer, occasionally with the help of one or both of the two other authors.

Each experiment lasted from 120 to 131 daily sessions. For most sessions, the 80 pictures were drawn randomly from a pool of about 500–700 (half positive, half negative). Consequently, some pictures were being seen for the first time, and others were repeaters, with the proportion shifting toward repeaters as the experiment progressed. On certain sessions, however, only new pictures were presented, and it is the data from these that have been analyzed here. Unless otherwise stated under Results, these pictures were distinct in no way other than being shown for the first time. No effort was made to improve learning by using easier pictures early in training or by any other explicit sequence.

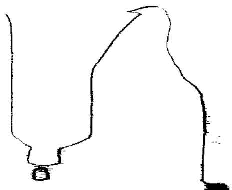
Results

Discrimination

Data from three experiments are presented together to aid comparison. Each session yielded a rate of pecking for each of the 80 pictures. Since the rate of pecking may be associated with the prior occurrence of reinforcement during the 10–90 sec that a picture is presented, the responding during a given positive stimulus after the first reinforcement is suspect. Figure 1 consequently shows the "corrected" rate of responding during positive stimuli, which means just the responding prior to the first reinforcement for each positive stimulus. During negative stimuli, a different need for correction existed. Negative stimuli cannot end within 10 sec of a peck. This penalty for pecking can artifactually depress the apparent rate of pecking, for it keeps the stimulus on until the response rates gets low. The corrected rate of negative-stimulus responding is for the responding during the scheduled 10–90 sec of presentation, that is, prior to the imposing of any penalty for responding.

The data in Figure 1 show 1 out of 10 consecutive sessions using only new pictures (i.e., 800 new pictures per experiment), on which most of this analysis of results is based for nine of the 11 subjects. These 10 sessions were run after about 75 sessions using pictures from the general pool. For all pigeons, including the two omitted in Figure 1, the positive pictures occasioned higher average rates of responding than the negative. Each pigeon also erred occasionally, responding slowly to positive stimuli or rapidly to negative. Beyond these generalities, not much obvious uniformity is displayed here by the subjects in each experiment. The absolute rates of responding varied widely, as shown by the adjustments in the scale of the y axis. Taking the subjects individually, rates of responding drifted up, down, or otherwise fluctuated during the session.

To assess statistical significance, the Mann–Whitney *U* test appears to be appropriately conservative (Siegel, 1956). If discrimination were perfect, the 40 highest corrected rates of responding would be to the positive stimuli, whereas the 40 lowest would be to the negative (assuming a 40/40 split). The absence of discrimination would result in a random mixture of positive and



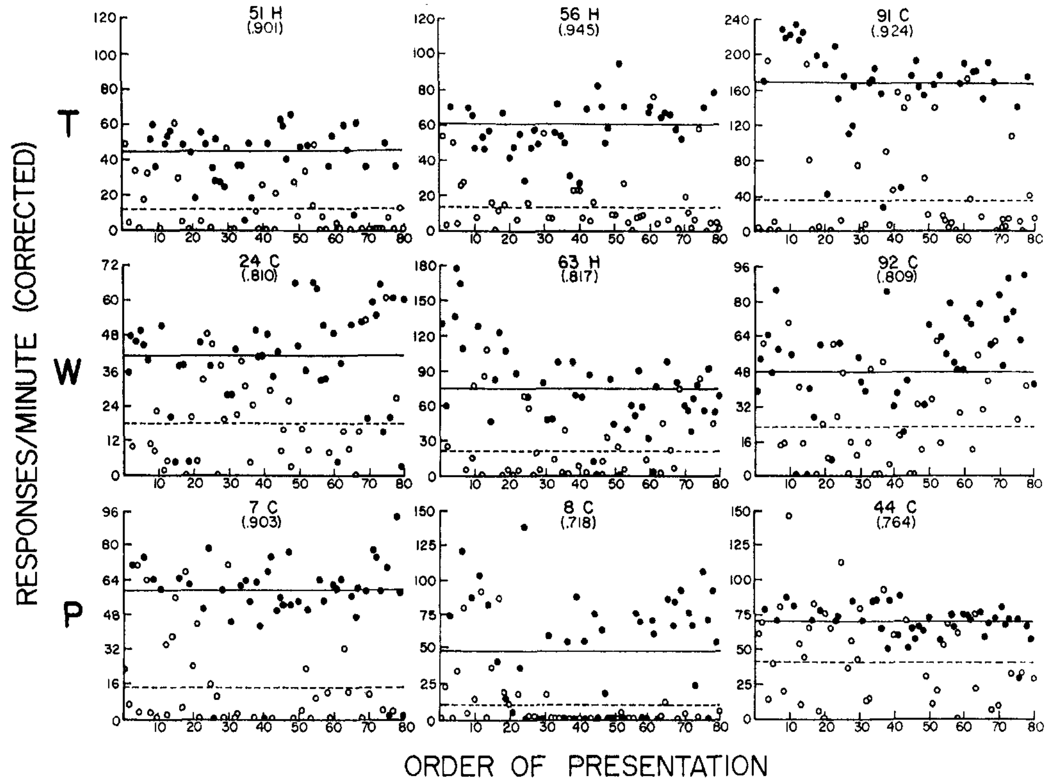


FIGURE 1. Corrected rates of pecking in the presence of each picture during a representative session using only new pictures, plotted against the order in which the pictures were shown. (Each coordinate axis is for one subject; each row for one of the three experiments. Filled circles are for positive stimuli; open circles, for negative stimuli. The solid horizontal line averages the filled circles; the dashed line averages the open ones. The decimals in parentheses are values of an index of discrimination, see text.)

negative stimuli throughout the list of ranks. *U*, which measures the degree of nonrandomness in the obtained ranking, confirmed the presence of discrimination in all nine cases in Figure 1. Discrimination was weakest for 8C, but the probability of drawing the obtained rankings by chance was nevertheless well below .01 even in this case. For the sessions shown in Figure 1, discrimination was strongest for 56H, where the associated probability was below .000,001.

Altogether, 108 sessions of data for new pictures were obtained for the 11 pigeons during the 10 days of tests (not 110, for there were two data-recording failures). The *U* test failed to achieve, or marginally achieved, a significance level of .05 five times, spread over the three experiments. The remaining 103 sessions were well, often far, into the range of statistical significance. Combining the tests of significance produced probability values so infinitesimally small that they are usually not tabled.

Although the *U* test establishes nonrandom ranking, it is not conveniently interpreted as a measure of the precise degree of nonrandomness. For that purpose, we divided the value of *U* for each session by the product of the numbers of positive and negative pictures to get an index ρ, a proportion that estimates the probability that the rank for the responding to a positive stimulus is above that to a negative stimulus (Bamber, 1975; Bradley, 1968). Bamber (1975) has shown that this quantity equals the area below a receiver operating characteristic graph whose coordinates are the probability of a positive instance being above a given rank and the probability of a negative instance being above that rank. When discrimination is perfect, ρ should be 1.0; when it is absent, it should be .5. Figure 2 shows the obtained values of ρ for the 10 sessions of testing with new pictures.

Taking the median ρ values (see Figure 2) as the measure of discrimination, pictures containing trees have about a 90% chance of eliciting higher pecking rates than pictures without trees for three of the four pigeons in Experiment T. Pigeon 244C, the fourth pigeon, had a median ρ of .745. The average of the four medians was .853. This is higher than the average of the medians in Experiment W, which was .790, or the average of the medians in Experiment P, which was also .790. Experiment T thus produced a slightly more accurate discrimination than the other two, but all three experiments demonstrated a concept involving a class of natural objects (i.e., pictures of those objects). The general level of discrimination in Figure 2 was no different from that shown in the training sessions using a mixture of old and new stimuli.

Comparisons across the three experiments are not like comparisons of discriminability in experiments using fixed stimuli in which a poor score means a poor discrimination. Here, the degree of discriminability must reflect to some extent the kind of exemplars we choose for a class. A poor score with obscure pictures does not signify that the discrimination has been poorly learned. We cannot, therefore, infer that trees are even slightly more discriminable than the other two classes, for the difference may have been in the selection of instances.

Figure 2 contains bands of cross-hatching on each coordinate. These show the region[1] of values of ρ that would have been obtained if the Mann-Whitney U had been statistically insignificant at the .05 level. Values of ρ ±.1 of .5 indicate insignificant nonrandomness in the rank order of rates of responding. As noted earlier, this occurred no more than five times in the 108 tests. Figure 2 shows that four of those five were just marginally insignificant.

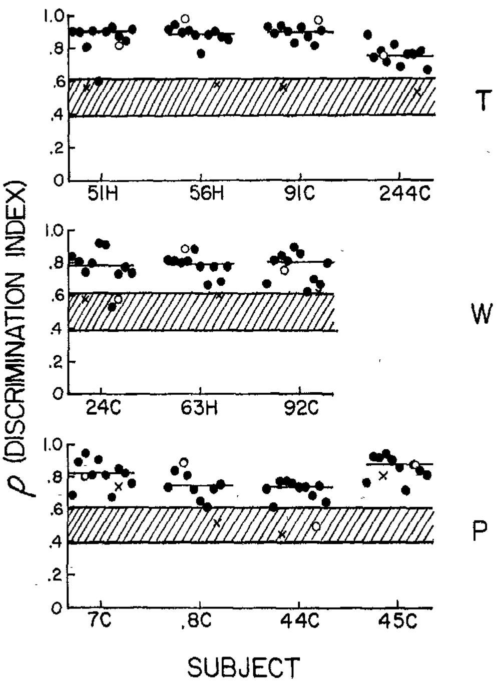


FIGURE 2. Discrimination index, estimating the probability that a positive stimulus will be ranked higher than a negative stimulus. (Each filled point gives the value of ρ for a single session using new pictures. The short horizontal lines show the medians of ρ for individual subjects. The cross-hatched region shows where values of ρ are statistically insignificant. Open circles and Xs are for special sessions, see text.)

Without seeing the actual pictures used, it is hard to grasp the range of stimuli handled by the subjects in the experiment. Figures 3–5 contain representative stimuli for each experiment in black and white instead of the original color. The pictures were chosen from among those that no more than one pigeon misclassified, and in many cases no pigeon misclassified, on first viewing. The pictures in Figures 3–5 were chosen to suggest the variety of stimuli, although four pictures per experiment barely

[1] The small daily variation in the numbers of positive and negative stimuli causes a small variation in the region of statistical insignificance. We have omitted this complication in Figure 2 because it would have added nothing to the data.

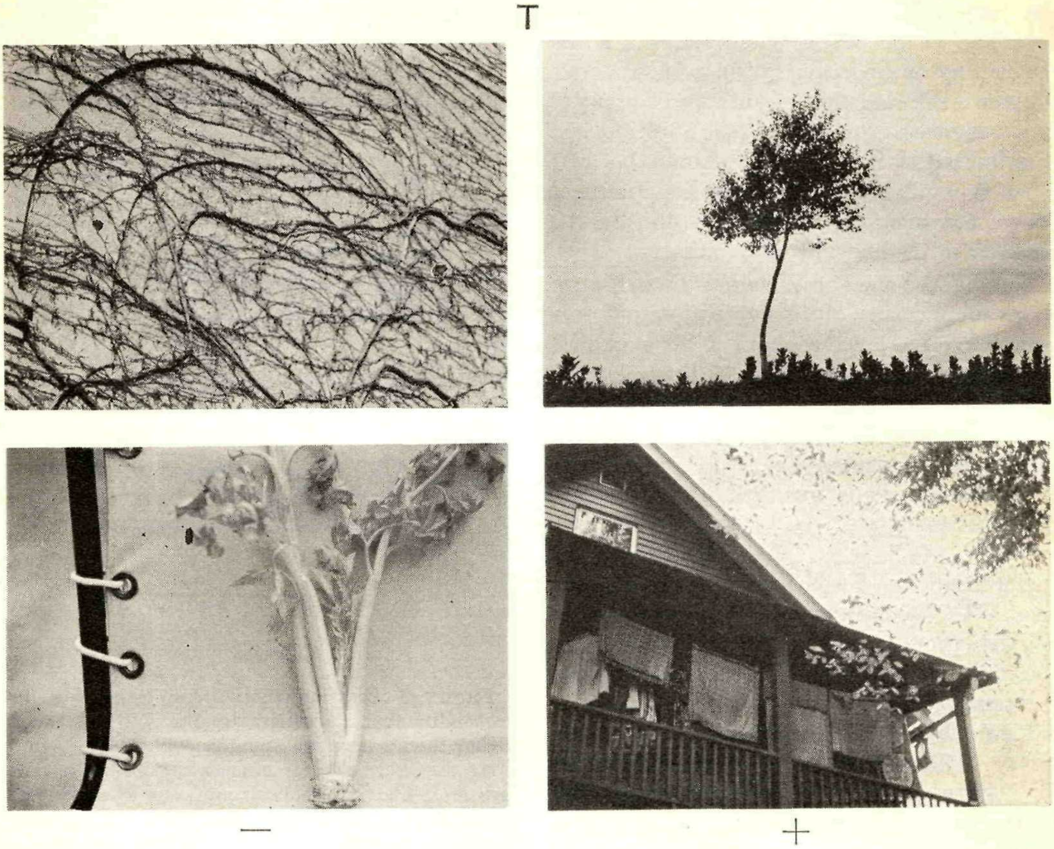


FIGURE 3. Four typical pictures used in Experiment T (trees), which were correctly classified by at least three of the four pigeons. (Negative stimuli are on the left; positive, on the right. The upper left picture shows a vine climbing on a cement wall; the lower left, celery.)

begin to do justice to the hundreds of pictures actually used.

In summary, from aggregate measures of performance and from an inspection of the pictures themselves, it is clear that the pigeons used principles of classification that approximate those we use ourselves, at least in complexity.

Concordance

Statistically, the problem of concordance resembles the problem of agreement among a group of judges at, for example, a flower show. Each judge ranks all the entries independently of the other judges. For any pair of judges, agreement may be evaluated with a measure such as Spearman's rank correlation. For more than two judges, a convenient measure is Kendall's coefficient of concordance, W (Hays, 1963; Kendall, 1948; Siegel, 1956), which is closely related to the average of the Spearman rank correlations between all pairs of judges. Unlike correlation coefficients, W falls between 0 and 1.0, with complete concordance equal to 1.0. For sizable (> 7) numbers of items to be ranked (n) the statistical significance of W is approximated by the chi-square distribution with $n - 1$ degrees of freedom.

For each session of new pictures, two coefficients of concordance were calculated, one for the ranked rates of responding to positive stimuli and one for that to negative stimuli. We were thus assessing the *degree* to which the pigeons agree, above and beyond their agreement about which pictures were positive or negative. Had W

been calculated for the entire day's session, then it would have reflected concordance owing to the discrimination between positive and negative stimuli. Any concordance in our analysis, however, is not explained by the separation of the ranks of positive and negative stimuli.

Table 1 gives the values of *W* for the 10 test days per experiment, where possible. Owing to several minor imperfections in the record of individual subjects, 3 days of the 30 are missing. The probability level is generally a function of *W*, but not exactly, for the chi-square corresponding to each value of *W* depends on both the number of pictures and the number of subjects, both of which varied to some extent. Nevertheless, a clear pattern emerges. Intersubject agreement is consistently and markedly greater for negative stimuli than for positive. Assuming no artifacts, concordance can only be shown when the pictures in a group vary in discriminability for the subjects individually. That is to say, each subject must find at least some pictures harder to classify than others. Then, concordance shows up when the subjects agree to some extent about which are hard and which are easy (or not so hard). If each subject finds all the pictures equally hard or equally easy, then concordance must be absent, for concordance is, like any correlation coefficient, a measure of covariance. The greater concordance for negative pictures may, then, be as much a reflection of greater negative picture variance as of anything else.

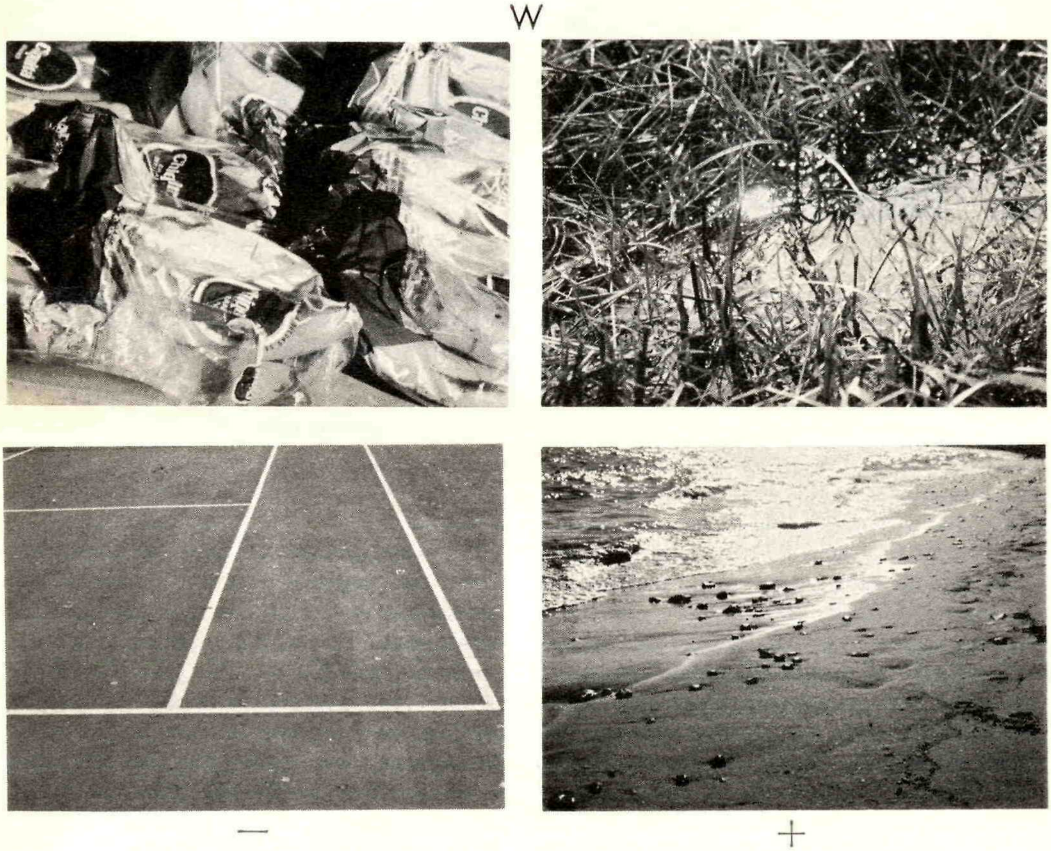


FIGURE 4. Four typical pictures used in Experiment W (water), which were correctly classified by at least two of the three pigeons. (Negative stimuli are on the left; positive on the right. The upper left picture shows cellophane bags of bananas; the upper right shows a small puddle through vegetation.)

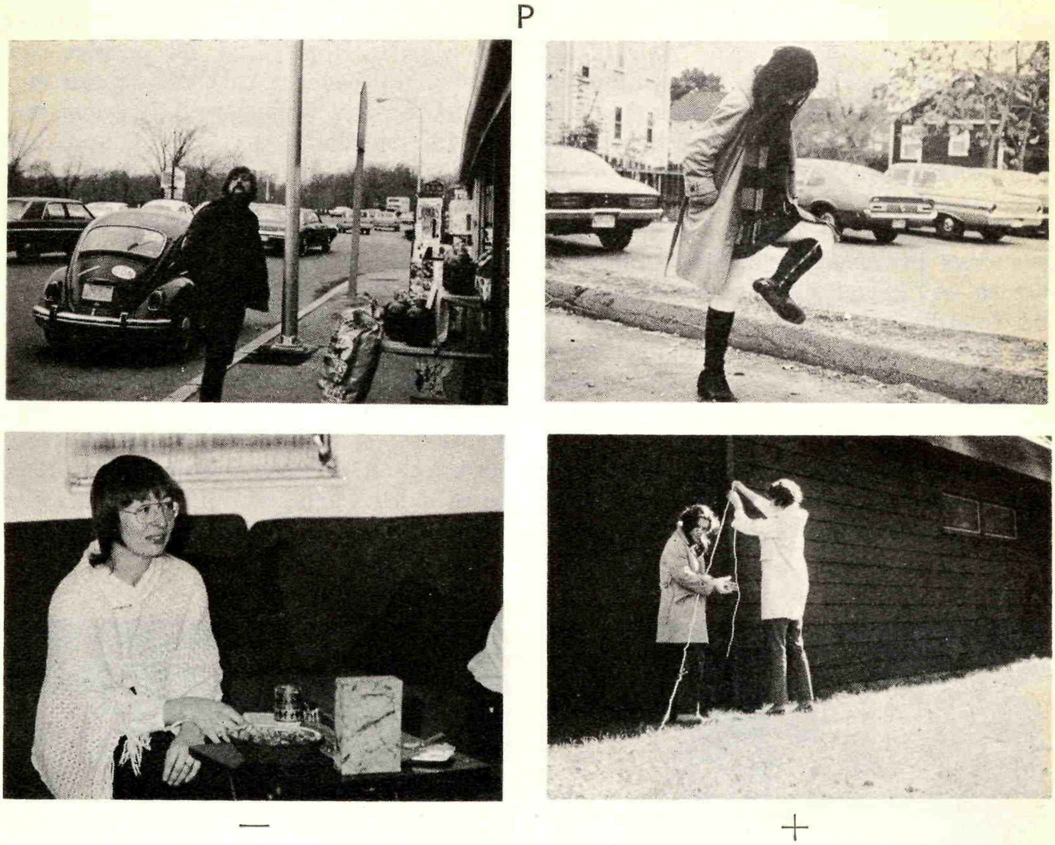


FIGURE 5. Four typical pictures used in Experiment P (specific person), which were correctly classified by at least three of the four pigeons. (Negative stimuli are on the left; positive, on the right. The upper left picture shows the subject's husband wearing her scarf; the lower left shows a different woman in the subject's apartment.)

This hypothesis can be tested by exploiting a feature of our procedure. Consider a session containing 40 positive and 40 negative stimuli. Perfect discrimination would be when all the rates to the positive ranked above all the rates to the negative stimuli. An error consists of ranking a positive stimulus among the negative stimuli, someplace among the bottom 40 ranks. However, this error will necessarily move one negative stimulus rate up among the top 40 ranks (assuming some negative-stimulus rates above 0). There is, in short, a corresponding negative misclassification for every positive misclassification and vice versa, a "true" error and, as a by-product, a "displacement" error. Greater negative-stimulus variance reduces to the hypothesis that true errors are more likely to be to negative stimuli than to positive, which is to say, to be false alarms rather than false dismissals.

The distributions of ranks of errors confirm that negative stimuli produced the greater number of true errors. Displacement errors should crowd the midrank, since they are being pushed down (for positive stimuli) or up (for negative stimuli) by the intrusions into the correct half of the ranks. In contrast, true errors may be distributed in any pattern within the incorrect half of ranks, depending on the degree of certainty in the erroneous ranking. Figure 6 presents the relevant analysis. Data were pooled across all subjects in each experiment and across the 10 sessions used for the coeffi-

TABLE 1

Coefficients of Concordance, *W*

	Trees[a]		Water[b]		Person[a]	
Session	Positive	Negative	Positive	Negative	Positive	Negative
1			.49**	.54***	.41***	.55****
2	.27*	.47****	.40*	.66****	.38**	.40***
3	.36**	.46****	.43*	.63****	.36**	.44****
4	.22*	.47****	.53**	.69****	.52****	.54****
5	.28*	.31*	.41*	.52**	.39**	.48****
6	.55****	.55****	.41*	.55***	.27*	.48****
7	.25*	.39**	.60****	.55***	.27*	.59****
8	.38**	.46****	.38*	.63****	.21*	.63****
9	.57****	.33**	.46**	.63****		
10			.33*	.54***	.32*	.54****

[a] $n = 4$.
[b] $n = 3$.
* $p > .10$.
** $.01 < p \leq .10$.
*** $.005 < p \leq .01$.
**** $p \leq .005$.

cients of concordance in Table 1. The abscissa gives the incorrect half of ranks ("worst 40") in blocks of 10. For the positive stimuli, this is the 40 lowest ranks of rate of responding; for the negative stimuli, it is the highest 40. To make them directly comparable, the abscissa goes from the midrank to the end for both functions. That is to say, the block called 50 comprises the 41st–50th rank for positive and negative stimuli, but the ranking went from high to low rates for the positive stimuli and from low to high for the negative, and so on.

The ordinate shows the proportion of negative or the proportion of positive stimuli contained in the corresponding block. For example, in Experiment T, .112 of all positive stimuli were between the 41st and 50th rank, but only .090 of all negative stimuli were. The block called 80 shows what proportion of all positive stimuli were in the lowest 10 rates of responding and what proportion of all negative stimuli were in the highest 10 rates of responding. For Experiment T, the figures were .023 and .039, respectively.

With equal numbers of positive and negative stimuli, the four values plotted for the two curves for each experiment in Figure 6 would have summed to equality.[2] (Figure 6 shows in decimals what the actual total proportions were, reflecting the slight excess of negative stimuli.) In each experiment, the curve for positive stimuli is more sharply decreasing than that for negative stimuli, thereby showing more crowding below the midrank. This indicates that there were more displacement errors for positive stimuli than for negative or, inversely, that there were more true errors for negative stimuli than for positive. We may therefore infer that the subjects were more prone toward false alarms than false dismissals. Whatever its origin, the asymmetry may account for the greater concordance among negative stimuli.

A control session eliminated one possible source of the greater negative-stimulus variance in ranks, based on our method for correcting rates of responding. For positive stimuli, we used response rate prior to the first reinforcement; for negative stimuli, response rate prior to the penalty for responding (see section above entitled Discrimination). Consequently, the average durations for positive and negative samples differed, although the distributions overlapped. The difference between positive and negative

[2] When there are 40 positive and 40 negative stimuli, the proportion of correct and incorrect instances can be represented as: $P_C + P_I = 1.0$, and $N_C + N_I = 1.0$. The errors comprise P_I, the proportion of positive stimuli ranked below the midline, and N_I, the negative proportion above the midline. With equal numbers of positive and negative stimuli, the proportions in the upper 40 ranks (for the 40 highest rates) must be: $P_C + N_I = 1.0$. By substitution, it follows that $P_I = N_I$.

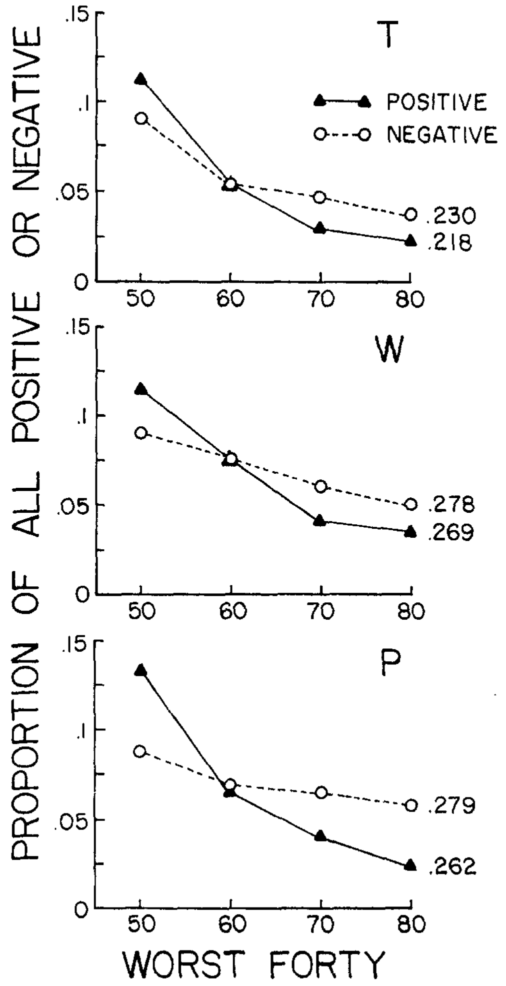


FIGURE 6. Proportions of positive stimuli ranked below the midrank and of negative stimuli ranked above the midrank in each experiment, averaged across subjects. (The "worst 40" refers to the lower half of ranks for positive stimuli and the higher half for negative stimuli. In blocks of 10 ranks starting at the midline, functions show the proportions of positive and negative stimuli. Total proportions among the worst 40 are given by the decimals on each function. Positive-stimulus functions include the false dismissals; negative-stimulus functions, the false alarms.)

samples ranged around 7 to 12 sec; in proportional terms, the negative samples were anywhere from 20% to 50% longer. It is possible that the longer negative stimulus durations allowed more concordance to develop, which might have accounted for the differences in Table 1 and Figure 6.

To check this possibility, the seven subjects in Experiments T and W were run with an additional session of new pictures and a modified definition of the corrected rate for negative stimuli. The variable-interval schedule for food reinforcement was run in the computer during both positive and negative stimuli. During positive stimuli, it continued to program reinforcements as before. During negative stimuli, it timed the interval over which rate of responding was calculated, so that the peck that would have been reinforced had the stimulus been positive shuts off the response-rate recorder. The average durations for the corrected rates to positive and negative stimuli for the seven subjects on the additional test day had a mean difference of .7 sec, with positive durations longer.[8]

The question is whether concordance still favors negative stimuli, even when the difference in durations is eliminated. For Experiment T, the negative stimulus concordance was .44 and the positive stimulus concordance was .32. Using the chi-square approximation, the negative stimulus value was significant beyond the .005 level, and the positive stimulus value was insignificant at the .10 level. For Experiment W, the negative stimulus concordance was .52 and the positive stimulus concordance was .42. This positive stimulus value is again insignificant at the .10 level, and the negative stimulus value has an associated probability level between .10 and .01. From this, we can conclude that negative stimulus concordance is higher than positive even when the response-rate artifact has been eliminated.

A second source of the findings in Table 1 and Figure 6 has not been definitively excluded. All subjects in each experiment saw stimuli in the same order. Concordance may

[8] This modified definition of corrected rate was used for all sessions and all subjects in the special sample, to be discussed, which demonstrates that discrimination itself was not dependent on the definition of corrected rate. In general, we found no evidence that this change made any difference. Indeed, from casual inspection it seems that much the same picture would have emerged had we not corrected rates at all, but we have not redone our entire analysis with uncorrected rates.

therefore reflect systematic variations in pecking during each session, which could differ for positive and negative stimuli. A sampling of product-moment correlations between rank and order-of-presentation number are shown in Table 2, in which the session numbers correspond to those in Table 1. Briefly, although most of the correlations are small, there is clear evidence for a trend toward lower ranks later in the session (shown by negative correlations), such as may be produced by food satiation, for example. Even though these correlations doubtless contribute to the concordance among subjects, it does not seem possible to specify the size of the contribution or to relate it to the difference between positive and negative stimuli given the present results.

For purposes of examining correlated errors, we identified the eight worst positive and eight worst negative stimuli in each of the 10 sessions of new stimuli used in this analysis. That is to say, we retrieved for each subject in each session the eight positive stimuli that obtained the lowest ranks in rate of responding and the eight negative stimuli that obtained the highest ranks. Given the contingencies of the experiments, these stimuli must be taken as worst in the sense that they come closest to misclassification; hence, they are called the "worst cases" in the ensuing analysis.

The pictures shown earlier in Figures 3–5 were considered correctly classified because they fell into these eight worst cases for no more than one subject in each experiment. Here we focus on those that fell among the worst cases for several subjects. Figures 7–9 show two positive and two negative pictures for each experiment that at least $n - 1$ pigeons misclassified, again in black and white instead of color. The pictures in Figures 7–9 are mostly, though not invariably, relatively difficult for human observers. Even if errors were uncorrelated, however, some pictures would fall among the worst cases for $n - 1$ subjects by chance. The preponderance of difficult stimuli in Figures 7–9 is consistent with the conclusion that the subjects' rule for classification approximated the experimenters'.

TABLE 2

CORRELATIONS BETWEEN RANK OF RESPONDING AND ORDER OF PRESENTATION

	Session 2		Session 8	
Subject	+	−	+	−
	Trees			
51H	.04	−.22	−.15	−.35
56H	.43	−.19	−.05	−.20
91C	−.26	.06	−.34	−.23
244C	−.27	−.16	−.15	−.45
	Water			
24C	.09	.23	.47	.02
63H	−.42	.04	−.61	−.45
92C	.35	.10	−.09	−.35
	Person			
7C	−.12	−.29	−.34	−.37
8C	.10	−.38	−.08	−.37
44C	−.32	−.24	−.30	−.06
45C	−.07	−.36	−.06	−.31

Special Samples

The data so far summarized do not prove that the subjects' categories were isomorphic with the experimenters', though they suggest it strongly. It remains possible that an incidental flicker of light, a stray sound of switching circuitry, or something comparable was correlated with positive or negative stimuli and controlled performance without our knowledge. Although patent artifacts were deliberately guarded against, more subtle ones may not have been anticipated. As a control for stimulus artifacts, two sets of 80 pictures for each experiment were handpicked. One set of pictures was picked to be easy and the other hard by the experimenters' own standard of judgment. The range of difference was roughly comparable to the easy and hard stimuli in Figures 3–5, but the pictures actually used were from among those never previously shown to the pigeons. Then, on 2 consecutive days, each pigeon saw first one, then the other special samples. Some pigeons saw the "hard" pictures the first day and the "easy" pictures the next; other pigeons had the reverse order. The procedure was the same as for the 10 test sessions described earlier, except for the calculation of corrected rates (see

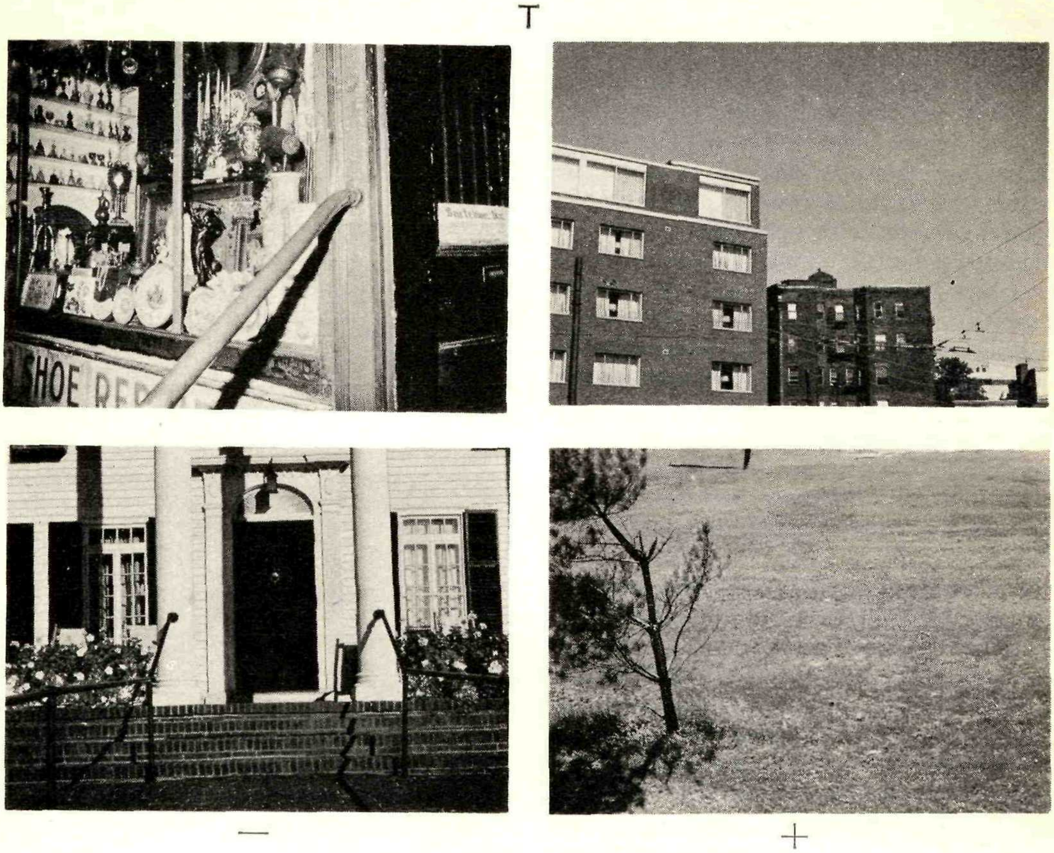


FIGURE 7. Four pictures misclassified by at least three of the four pigeons in Experiment T. (Negative stimuli are on the left; positive, on the right. In the upper right picture, trees are just barely visible.)

Footnote 3). Table 3 presents the main results of this special procedure in terms of ρ. The values of ρ also appear in Figure 2 as crosses (hard pictures) and open circles (easy pictures), in the order in which they were obtained.

If the pigeons' stimulus classes were essentially isomorphic with ours, and if ρ can be accepted as a valid measure of discrimination, then the expected pattern would be that shown by Subjects 56H, 91C, 63H, and 8C, for whom the easy pictures produced relatively high values of ρ while the hard pictures produced low ones. The five other pigeons deviated from the ideal pattern. Subjects 51H, 92C, and 244C found the hard pictures hard, but the easy pictures were no easier than the unselected pictures in the earlier tests. Subjects 24C and 44C, who found both hard and easy pictures hard, had seen the hard pictures first. The poor performance with the easy pictures may show the aftereffects of the hard pictures. Finally, Subjects 7C and 45C found the hard pictures only slightly harder than the easy ones, with both sets well within the range observed with unselected pictures.

As a whole, hard pictures had the more consistent effects. For 9 of the 11 subjects, discrimination fell to virtual statistical insignificance. For the other 2 subjects, discrimination was below the earlier median level but not markedly. Easy pictures produced a more complex mixture of effects. Five subjects discriminated them about as well as the unselected pictures. For 4 subjects, discrimination was what might be called supernormal. And discrimination was

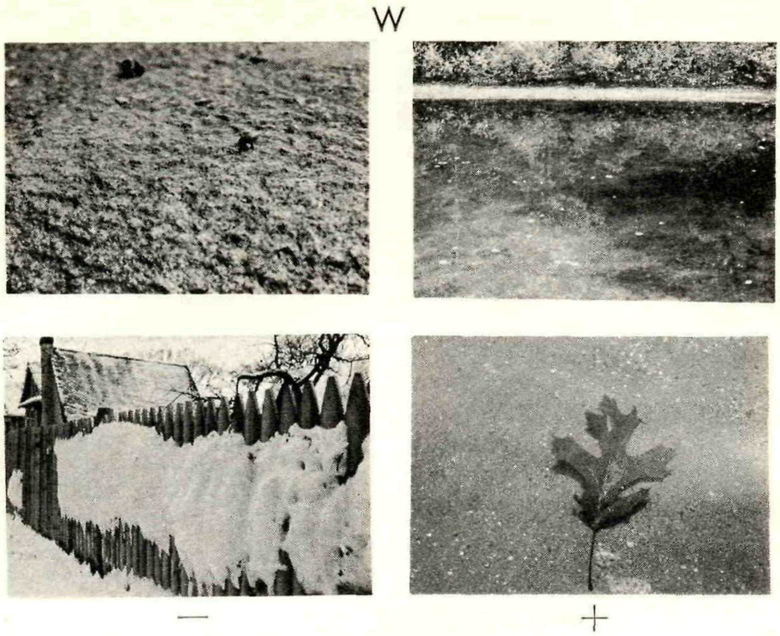


FIGURE 8. Four pictures misclassified by at least two of the three pigeons in Experiment W. (The upper right picture shows a swampy field in front of a road; the lower right picture shows a leaf under water on a sandy bottom near the shoreline.)

absent for 2 subjects that saw the easy pictures after having been disrupted by the hard ones. The overall effects of the hard and easy samples should dispel any doubt that the controlling stimuli were visual for all the subjects except possibly 7C, which showed no effects of either hard or easy stimuli. Moreover, it seems that the visual classes controlling performance were at least correlated with what we ourselves are looking at, though perhaps imperfectly. The test with special selections uncovered differences among the subjects that were not evident in their performances with hundreds of unselected pictures.

Discussion

The ability to discriminate open-ended classes of stimuli poses problems at two levels of analysis. First is the analysis of the features enabling a subject to tell whether an object is a member of a class—whether a picture contains a tree, for example. Second is the analysis of the properties of classes that render them discriminable. Let us assume that the thousands of pictures used in the present experiment could have been divided into some sets that pigeons could, and some that they could not, learn to sort. Above and beyond the question of how they sort within any given classification problem is the question of distinguishing between solved and unsolved problems.

The traditional explanation at the first level of analysis is a theory based on common elements, recently exemplified by Blough (1975) and Rescorla (1976). Trees, according to this theory, have something specific in common, for example, a certain shape or texture or color or combination of them. And if common elements suffice for the first level, they also solve the problem of the second level. A classification would presumably be discriminable only if it were based on discriminable common elements. However, having looked at the hundreds of instances used here or even at the two positive pictures in Figure 3 (let alone the tens of thousands involved in real-life discriminations), we cannot begin to draw up a list of

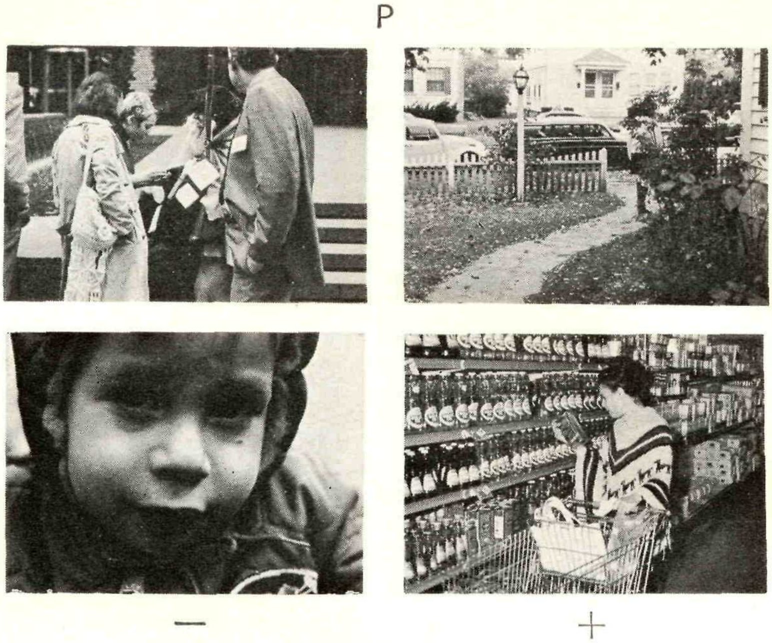


FIGURE 9. Four pictures misclassified by at least three of the four pigeons in Experiment P. (The lower left picture shows the subject's child, who occasionally appears with her in positive stimuli.)

common elements. To recognize a tree, the pigeons did not require that it be green, leafy, vertical, woody, branching, and so on (overlooking the problem of common elements nested within terms like leafy, vertical, woody, and so on). Moreover, to be recognizable as a nontree, a picture did not have to omit greenness, woodiness, branchiness, verticality, and so on. Neither could we identify common elements in the other two experiments.

TABLE 3

DISCRIMINATION INDEX, ρ, WITH SPECIAL SAMPLES OF STIMULI

Subject	Easy	Hard
	Trees	
51H	*.814*	.554
56H	.980	*.573*
91C	*.963*	.551
244C	.742	*.523*
	Water	
24C	*.574*	.578
63H	.881	*.590*
92C	.744	*.616*
	Person	
7C	.799	*.748*
8C	.886	*.514*
44C	*.497*	.450
45C	*.883*	.808

Note. Italicized indexes show the second of the two sessions of tests with the special samples.

If not common elements, what? No other theory is so easily characterized, though in crude terms an alternative suggests itself. Pigeons respond to clusters of features more or less isomorphic with the clusters we respond to ourselves. The green should be on the leaves, if either green or leaves are present. However, neither is necessary or sufficient. The vertical or branching parts should be the woody parts, although neither of these features is necessary or sufficient either. What we see as trees comprises a complex list of probabilistic conjunctions and disjunctions, the discovery of which would require far more effort than seems justified by any possible benefit. Insofar as no visual element or configuration of them is either necessary or sufficient, there can be no single prototype or schema defined at the level of visual arrays, a conclusion much like Wittgenstein's notion of "family resemblance" (Wittgenstein, 1953).

After a few dozen daily sessions amounting to no more than about 700 different instances of trees and nontrees, the pigeons readily sorted another 800 instances, none of which was identical to each other or to the original 700. New pictures were often discriminated with higher accuracy than pictures already seen. Even more impressive generalization was shown by pigeons in a study using pictures of people as the positives class (Malott & Siddall, 1972), in which just one or two dozen pictures were used in the original training. From this scanty exposure, the subject generalized to new instances without apparent limit. The minimal case appears to be Cerella's (Note 1) experiment, in which pigeons may have generalized to the class of silhouettes of oak leaves from having been trained with 1 positive instance and 40 negative instances.

These experiments could be considered cases of stimulus generalization. But unlike the typical experiment using standard physical dimensions as the stimulus variable, we know of no relevant physical variable. Instead of wavelength or decibels or seconds, the dimensions here can, at present, be characterized only in the language of objects, not stimulus attributes, a distinction embodied in Konorski's (1967) concept of the "gnostic field." The pigeons display semantic generalization in the sense that we can describe their behavior better by noting what the pictures are pictures *of*, rather than by what the pictures themselves *are*. Having seen a collection of patterns that we recognize as pictures of trees, the pigeons generalize to novel patterns that we also recognize as pictures of trees. The constancy is in the world of distal objects, not in that of proximal patterns.

Little is resolved, however, by saying that the performance is a form of generalization, semantic or otherwise, for we have no satisfactory theory of generalization either. The semantic categories used by the pigeons must have physical specifications, for the pictures are, in fact, nothing but optical arrays. They apparently see a stalk of celery, leaves and all, as a nontree, although it is green, leafy, etc. On the standard optical continua, the celery is clearly close to many of the trees one sees, but the pigeons shift to some other system of classification when it pays to do so, as it does here. Data from traditional generalization experiments show that pigeons are sometimes able to operate along the standard visual attributes such as color, shape, orientation, etc., about as well as we do, but, like us, they have the capacity to deal with object categories too.

The capacity doubtless has something to do with evolution. To the question of what distinguishes discriminable stimulus classes from the indiscriminable, a common response brings in the creature's germ plasm. It is held that organisms are disposed to group together those stimuli that signify objects with similar psychological consequences. For example, trees, bodies of water, and people have long been both important and common in the pigeon's natural environment. By now, these objects may have had enough evolutionary significance to be somehow represented in the genes. The evolutionary account is used not only by psychologists confronted by the data on categorization and generalization but also by philosophers (e.g., Quine, 1969) discussing the deficiencies of traditional empiricism.

Assuming for the sake of argument that there are genetic constraints on the categories that creatures induce, it remains to be shown how those constraints operate functionally. For example, what is it about the stimuli in Experiment T that triggered the genetic predisposition to induce trees but in Experiment W to induce water? The triggering factor could not have been greenness, woodiness, and so forth—those elements whose inadequacy led us to the genetic hypothesis in the first place. The genetic hypothesis shifts the locus of the problem of classification but does not solve it.

Although we did not discover the optical details of the categories used by our pigeons, certain conclusions can be drawn. The present findings show that classes as complex as trees can be defined by static features. Successful discrimination of still photographs proves that a pigeon need only look at examples of a class in order to activate its rules. Since, in nature, interactions are with objects as a whole, it is not a foregone conclusion that two-dimensional projections could work. It is also not a foregone conclusion that movement could be omitted. It has been suggested that people unaccustomed to looking at pictures have trouble finding objects (Deregowski, 1972), but the subjects in such studies get less practice than did the pigeons in ours.

It is also not a foregone conclusion that pigeons and people generalize similarly, yet they clearly did to a degree. For some pigeons, our "easy" instances proved easy and our "hard" instances hard. The pigeons were housed for several years in a room with windows on the seventh floor of a building in a residential neighborhood. From the pigeons' home loft, trees are visible in the distance (> about 200 yards; 180 m); the only visible water is in the drinking cups and perhaps in the occasional mop pail. The pigeons never saw the photographic subject in Experiment P in person. Nevertheless, pigeons and people converge on similar, if not equivalent, categories. The results suggest that the pictures used as stimuli activate a category rather than define it and, second, that the activated category draws on something other than past experience.

This is not to postulate innate categories for pictures of trees, water, and persons. Although that may seem a possibility for Experiment T, the analogous conclusion for Experiment P is manifest nonsense—an innate category for recognizing a particular young woman living in Cambridge around 1970. Instead, the innate ingredient in these discriminations must operate less specifically. Given a finite set of varying stimuli, the pigeon activates a particular category out of the limitless number of categories more or less equally well determined by the same set. It is in this narrowing of the range of possible categories that innateness seems to be expressed.

Data on human concept formation have indicated that a set of varying stimuli may be remembered as the central tendency of the variations called the "prototype" or "schema" (Posner & Keele, 1968). Even if the person never sees the prototype itself, retention will be more enduring for the prototype than for any of the individual exemplars whose central tendency defines the prototype (Posner & Keele, 1970). We noted earlier the trouble with a prototype defined at the level of the stimuli. In addition, there is some question about the generality of the finding for human subjects.

Rosch (1973, 1975) showed that the inferred prototypes deviate from the simple central tendency of exemplars when the stimulus domains are more "natural." Thus, whereas Posner and Keele used random configurations of dots as the stimuli in their experiment, Rosch used colors and forms in hers. The colors and forms that were prototypic tended to be perceptually unitary shades of red, green, and so on, or simple geometrical figures like squares, even when the exemplars were biased away from these prototypes and even when the subjects were drawn from populations (e.g., young children or New Guineans) that have not learned words for them. For Rosch, the "central tendency" characterization of prototype is augmented into a "best" or "clearcase" characterization. Pure, deep red becomes a prototype not because it is at the central tendency but because it meets a certain requirement or set of requirements so well. But this raises yet another question, as to the source of these requirements. In her answer, Rosch contrasts "perceptually given" prototypes, such as red or square, with "semantic category" prototypes, such as *bird* or *chair*. For the former, the best case prototype is "physiological," which means innate. For the latter, it is cultural, hence empiricistic, according to Rosch. The best cases of *chair, bird,* and so on, develop through some sort of learning process, which includes at least a component of taking a central tendency, though perhaps with nonlinear weighting of attributes. A quite similar notion, couched in neurophysiological vocabulary, is Konorski's (1967) gnostic unit, a neural element that somehow registers a percept of a distal object rather than of the proximal stimulus.

Rosch distinguishes between prototypes based on innate constraints and those based on experiential averaging—for example, *red* on the one hand, and *bird* on the other. *Trees, bodies of water,* and *person,* being semantic, would be experiential in her theory. But it seems improbable, to say the least, that the same experience-averaging procedure—however complex—could be applied to our samples of positive and negative stimuli in each of the three experiments and come up with roughly the right (i.e., the human) prototype in each case. And it is still more implausible applied to the previously noted studies by Malott and Siddall (1972) and Cerella (Note 1). It is more plausible to conclude that pigeons tend innately to infer a tree category from instances of trees, a familiar-person category from varying instances, and so on, more or less as we do ourselves. The properties of the inferred class arise from the joint constraints imposed by the stimuli and by innate factors. If two organisms form equivalent classes from the same stimuli, it is because they share not only the stimuli but also the genetic constraints.

Our data support at least a partial isomorphy in the inferred classes between pigeons and people. The pigeons discriminated reasonably, and they were in some cases suitably helped or hindered by the special easy or hard samples. These results favor the hypothesis of isomorphy. On the other hand, discrimination was not perfect and the special samples were not invariably effective, especially the easy ones. These results may point to something less than total isomorphy between the categories of the people and the pigeons involved here. However, we cannot say whether the difference is experiential or innate, for it could be either or both.

The most surprising result here may be not that pigeons can use open-ended natural categories but that the three experiments produced data as similar as they did. The sheer levels of discrimination (Figure 2), the concordance among subjects (Table 1), and even the pattern of true errors and displacement errors (Figure 6) are almost duplicated in each experiment. If we had imposed rigid criteria in the selection of pictures, such invariances would perhaps be less unexpected. But in fact, we had no such criteria. Each experiment produced more false alarms than false dismissals, more concordance for negative instances, and most discriminations in the range of $\rho = .7–.9$. The first two of those findings are consistent with the hypothesis of a smaller dispersion of positive stimuli than negative, a plausible finding given the stimulus materials. However, at this point, it must remain a hypothesis, not a conclusion, for the stimulus metric is itself unknown. Positive and negative stimuli may also differ inasmuch as correct positive-stimulus behavior produced food, whereas correct negative-stimulus behavior avoided delay. Further study is needed to see how much our method itself constrained both the levels of discrimination and the patterns of errors and concordance.

REFERENCE NOTE

1. Cerella, J. *Concept formation in the pigeon.* Unpublished manuscript, Harvard University, 1975.

REFERENCES

Bamber, D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology,* 1975, *12,* 387–415.

Blough, D. S. Steady state data and a quantitative model of operant generalization and discrimination. *Journal of Experimental Psychology: Animal Behavior Processes,* 1975, *1,* 3–21.

Bradley, J. V. *Distribution-free statistical tests.* Englewood Cliffs, N.J.: Prentice-Hall, 1968.

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

Deregowski, J. B. Pictorial perception and culture. *Scientific American,* 1972, *227,* 82–88.

Hays, W. L. *Statistics for psychologists.* New York: Holt, Rinehart & Winston, 1963.

Herrnstein, R. J., & Loveland, D. H. Complex visual concept in the pigeon. *Science,* 1964, *146,* 549–551.

Hull, C. L. *Principles of behavior.* New York: Appleton-Century-Crofts, 1943.

Kendall, M. G. *Rank correlation methods.* London: Griffin, 1948.

Konorski, J. *Integrative activity of the brain.* Chicago: University of Chicago Press, 1967.

Malott, R. W., & Siddall, J. W. Acquisition of the people concept in pigeons. *Psychological Reports,* 1972, *31,* 3–13.

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.

Quine, W. V. O. Natural kinds. In W. V. O. Quine (Ed.), *Ontological relativity and other essays.* New York: Columbia University Press, 1969.

Rescorla, R. A. Stimulus generalization: Some predictions from a model of Pavlovian conditioning. *Journal of Experimental Psychology: Animal Behavior Processes,* 1976, *2,* 88–96.

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

Rosch, E. H. Universals and cultural specifics in human categorization. In R. W. Brislin, S. Bochner, & W. I. Lonner (Eds.), *Cross-cultural perspectives on learning.* New York: Wiley, 1975.

Siegel, S. *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill, 1956.

Skinner, B. F. *Science and human behavior.* New York: Macmillan, 1953.

Wittgenstein, L. *Philosophical investigations.* (G. E. M. Anscombe, trans.). New York: Macmillan, 1953.

(Received September 10, 1975; revision received February 19, 1976)