Journal of Experimental Psychology:
Learning, Memory, and Cognition
1995, Vol. 21, No. 3, 662–673

0278-7393/95/$3.00

Prototype Effects and Peak Shift in Categorization

I. P. L. McLaren, C. H. Bennett, T. Guttman-Nahir, K. Kim, and N. J. Mackintosh
University of Cambridge

People asked to categorize exemplars of 2 categories often respond more accurately to the prototypes of those categories than to other exemplars. The authors suggest that this prototype effect may often have been confounded with a *peak shift* as is observed when pigeons are trained to discriminate between two wavelengths (S+ = 550 nm and S− = 560 nm), and the peak of their postdiscrimination gradient lies at 540 or 530 nm rather than at 550 nm. Three experiments established that a similar peak shift can occur when people are asked to categorize 2 sets of stimuli, but the authors also provide evidence of a true prototype effect uncontaminated by any peak shift. These results appear to pose considerable problems for exemplar-based theories of categorization.

If people are required to classify two sets of stimuli, patterns, shapes, letter strings, and so forth into two different categories in which each set can be described as a series of variants on two prototypes, they will often, on test, classify the two unseen prototypes significantly more accurately than new exemplars of either category (e.g., Posner & Keele, 1968); also there are instances of better performance on the prototypes or stimuli close to them than on the actual stimuli on which participants were trained (e.g., Homa, Sterling, & Trepal, 1981; Medin, Altom, & Murphy, 1984; Medin & Schaffer, 1978; Whittlesea, 1987). Such prototype effects are of course consistent with a theory that suggests that our concepts are stored in terms of typical or prototypical members or instances (Rosch, 1978). However, it is now clear that other theories of concept learning are equally capable of predicting these results. Exemplar theories, which assume that information about each training exemplar is stored in memory and that test performance is determined by generalization from trained exemplars, can predict such effects because prototypes, as the central tendency of the set of trained exemplars, will receive more generalized strength than any other novel stimulus (Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1986, 1987, 1989, 1991). Connectionist theories, which assume that training exemplars are stored as patterns of weight changes across a set of elementary units, can equally predict prototype effects because the elements activated by presentation of a prototype will be precisely those in which weights have changed most (Knapp & Anderson, 1984; McClelland & Rumelhart, 1985; Shanks, 1991).

There is another quite different experimental paradigm in which participants respond more accurately to stimuli they have never seen before than to the stimuli on which they have actually been trained. This is the *peak shift*, first reported in pigeons by Hanson (1959). Hanson rewarded pigeons for pecking at a key light illuminated with light of 550 nm and alternated these reinforced trials with others in which the key light was 560 nm and pecking was not rewarded. On subsequent test trials, the pigeons pecked more rapidly at wavelengths of 530 and 540 nm than they did to the original S+, suggesting that they were classifying these novel stimuli as more positive than the reinforced stimulus (S+) on which they had been trained. Similar effects have been obtained with other animals and other stimulus dimensions (e.g., Rilling, 1977). One explanation of this peak shift is to conceptualize a series of stimuli varying along any given dimension as sets of overlapping elements. If neighboring wavelengths, for example, 550 and 560 nm, share elements in common, these common elements will be associated with both reward and nonreward and might in consequence acquire little associative strength (cf. Wagner, Logan, Haberlandt, & Price, 1968). Response to the S+, 550 nm, would largely be controlled by those elements it does not share with the S−, 560 nm, but because these will precisely be the elements S+ shares in common with 540 nm and because they will form a more significant component of the representation of the latter wavelength than of S+, it follows that responding to 540 nm, containing only elements associated with reward, might be stronger than responding to S+ (Blough, 1975).

Now consider the case of two categories, each defined by a prototype, and each having exemplars that are distortions of the category prototype. Some of these exemplars may be close to the boundary between the two categories, others will be distant from this boundary. More particularly, some exemplars will lie between the two prototypes (in some similarity space), whereas others will lie beyond the two prototypes. The former will be closer to the other category's prototype than its own prototype is, the latter will be further away.

In principle, an exemplar's distance from the prototype of the other category can vary even when its distance from its own prototype is fixed. Evidence of prototype effects is said to be provided when performance on a novel test exemplar is shown to be a function of its distance from the prototype of its own

I. P. L. McLaren, C. H. Bennett, T. Guttman-Nahir, K. Kim, and N. J. Mackintosh, Department of Experimental Psychology, University of Cambridge, Cambridge, England.

This research was supported by a grant from the Science and Engineering Research Council. We thank Bruce Whittlesea for his many helpful comments on this article.

Correspondence concerning this article should be addressed to I. P. L. McLaren, Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB England. Electronic mail may be sent via Internet to iplm2@cam.phx.

category. In practice, this is often confounded with the exemplar's distance from the prototype of the other category. For example, in Medin and Schaffer's (1978) experiments, *polymorphous* categories were used, defined by the presence or absence of certain features. In the case of such categories, the prototypes are the exemplars that possess all of the defining features for one category and none of the other. This makes it inevitable that the prototype of one category is also the member of that category furthest away from the prototype of the other, which in turn makes it difficult to decide whether any advantage in ease of classification enjoyed by the prototype is due to its relationship to its own category or the other one. This could be construed as a confound between a prototype effect and peak shift.

Our aim is to systematically investigate the roles of prototype effects and peak shift in categorization. To this end, we define two types of category exemplar. *Close* exemplars are those that are more similar to the prototype defining the other category than their own defining prototype is. The distance (in some similarity space) between a close exemplar and the prototype of the other category is smaller than that between the two prototypes. *Far* exemplars are the converse; they are less similar to the other category prototype than their defining prototype is to that other prototype. The distance between a far exemplar and the other category prototype is greater than that between the two prototypes, and so it is further away from the other category than its prototype. Now we can ask the question: How does distance from the other category prototype interact with an exemplar's distance from its own category prototype? We report three experiments using computer-generated checkerboard patterns that examined whether prototypes are classified more accurately than training exemplars and whether such novel far exemplars are classified at least as well as prototypes.

Experiment 1

Method

Participants and apparatus. The participants were 35 undergraduate and graduate students or junior research assistants at Cambridge University or the Medical Research Council Applied Psychology Unit, Cambridge. The experiment was run on an Apple Macintosh Plus computer. The stimuli were checkerboard patterns consisting of 16 × 16 arrays of black and white squares, each square 4 × 4 pixels in size. At the start of the experiment, the computer generated a particular random pattern of black and white squares for each participant, which served as that participant's first prototype. It then changed 112 of these 256 squares (7 in each row) from black to white or vice versa to generate the participant's second prototype. Of these 112 squares that differed between the two prototypes, a randomly chosen 25 were changed to generate each close exemplar. A new close exemplar was generated for each training trial. A final set of far exemplars was generated by changing 25 of the squares that the two prototypes shared in common. All of the changes were made without replacement; that is, if a square was changed, then it was ineligible for any subsequent change. Table 1 provides a summary description of the stimuli, and Table 2 provides a schematic illustration of the results of the algorithm used for generating them. P1 and P2 are simplified versions of the prototype stimuli used in this experiment. C1 and C2 are instances of close exemplars, and F1 and F2 are instances of far exemplars from both categories. The proportion of elements (i.e., squares) changed for these exemplars is as close as possible to that used in the experiment, given the constraint of only having 16 elements rather than 256. Note that there are 9 elements common to the prototypes (the first 9 columns) and that they differ on 7 (in the final 7 columns). Elements shown in bold are those changed to construct that exemplar from its prototype.

Procedure. The participants were seated in front of the computer and instructed that they were going to be shown a number of different checkerboard patterns; their task was to decide whether each pattern was an instance of Category 1, in which case they should press the *x* key, or of Category 2, in which case they should press the period key. They were told that each stimulus would appear after a brief warning stimulus (a + sign) and would stay on the screen for only 4 s. They were instructed to respond as accurately as possible within these time constraints and were also told that they would receive feedback on whether they had been correct immediately after their response. Once this had been explained to them, they were asked to obey the instruction on the screen "press space bar to begin," and training trials began. Each trial started with the presentation of the + sign in the center of the screen, followed 1 s later by a randomly chosen close exemplar from one or other category. This pattern stayed on the screen for a maximum of 4 s, or until the participant responded, when

Table 1
Algorithm for Generating the Prototypes and Exemplars for the Two Categories Used in Experiment 1

Characteristic	Prototype 1	Prototype 2
Design	Random	Seven squares per row of Prototype 1 changed
Exemplar		
Close	Change 25 squares of the 112 that differ between Prototypes 1 and 2	Change 25 squares of the 112 that differ between Prototypes 1 and 2
Far	Change 25 squares of the 144 that Prototypes 1 and 2 share in common	Change 25 squares of the 144 that Prototypes 1 and 2 share in common

Note. Within the constraints given, squares are chosen at random, and once changed are not changed back.

Table 2
Schematic Illustration of the Type of Stimuli Generated by the Algorithm in Table 1

Prototype or exemplar	Elements															
P1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
P2	1	0	1	0	1	0	1	0	1	1	0	1	0	1	0	1
C1	1	0	1	0	1	0	1	0	1	**1**	**0**	0	1	0	1	0
C2	1	0	1	0	1	0	1	0	1	1	0	**0**	**1**	1	0	1
F1	1	**1**	**0**	0	1	0	1	0	1	0	1	0	1	0	1	0
F2	1	0	1	**1**	**0**	0	1	0	1	1	0	1	0	1	0	1

Note. If 1 = black square and 0 = white square, then the vectors define 4 × 4 checkerboards, with approximately the correct distribution of common and distinctive elements. A 1 or 2 designates the two categories. Elements shown in bold are those changed to construct that exemplar from its prototype. P = prototype; C = close exemplar; F = far exemplar.

the pattern disappeared to be replaced for 1 s by the signal "correct" or "error" (plus on error trials only, a bleep). The next trial followed immediately. Other comments printed on the screen were "you pressed the wrong key," if the participant pressed neither the x nor the period key; "time out," if they failed to respond within 4 s of the pattern appearing on the screen; and "you've anticipated the probe," if they had responded before the pattern appeared.

The sequence of stimuli shown was random with the restriction that there be 20 of each category's exemplars in each block of 40 trials. Participants were given a minimum of 40 trials and a maximum of 100 trials to learn the discrimination. Training could stop at any time after 40 trials, when a participant's total proportion of correct responses, up to and including that trial, was significantly ($p = .05$, two-tailed) above chance. If participants failed to learn within 100 trials, they were dropped from the experiment. Ten participants were discarded for this reason, leaving 25 to continue onto the test trials.

At the end of training, the instruction "press space bar to continue" appeared on the screen and this response initiated two blocks of 40 test trials. Each block contained 16 trials with new close exemplars, 16 with far exemplars, and 8 with the two prototypes; half of each class of stimulus from each category. Within each block of five trials there were two close exemplars, two far exemplars, and one prototype trial. Correct and error signals were provided after each trial, exactly as in training, and also as in training, each exemplar (close or far) was randomly generated for each trial; only the prototypes were repeated. At the end of the first 40-trial block of testing, the signal "press space bar to continue" appeared on the screen again to allow participants a brief rest before the second block of test trials.

Results

Twenty-five participants successfully learned the initial discrimination. Two measures of performance on test trials were taken: percentage correct and reaction time (on correct trials only). The results of these two measures are shown in Figure 1 for the three types of test trial. It is evident that participants responded both more rapidly and more accurately to the far exemplars than to the close exemplars but that performance

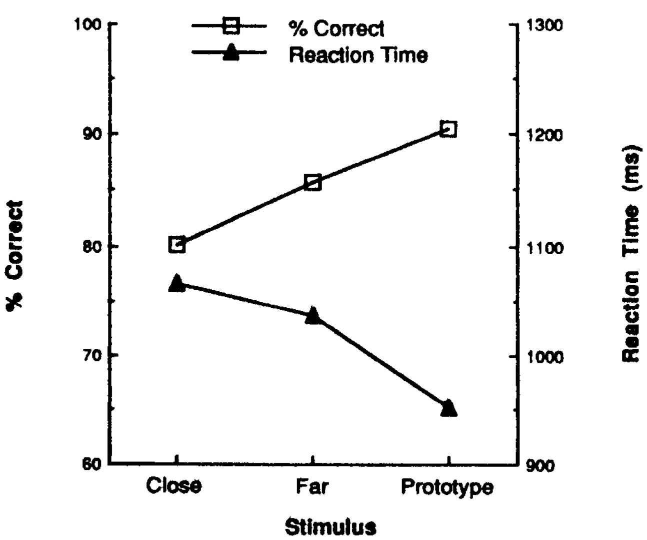

Figure 1. The results for Experiment 1 on both percentage correct and reaction time measures.

was most rapid and accurate to the two prototypes. An analysis of variance (ANOVA) indicated significant differences between the three types of trials, both on the percentage correct and on the reaction time measures, $F(2, 48) = 10.33$ and 16.67, $p < .01$, $MSE = 65$ and $5{,}373$, respectively. Subsequent Newman–Keuls tests established that performance on prototypes was both faster and more accurate than performance on both close and far exemplars, and performance on far exemplars was more accurate but not significantly faster than on close exemplars (all at least $p < .05$).

These test results suggest that not all exemplars are equally well classified: The close exemplars, even though necessarily more similar (by design) to the training stimuli than the far exemplars and no further from the prototypes, were less accurately classified than the far exemplars. This suggests that distance from, or similarity to, the other category is having the effect suggested by the peak-shift data. However, it is equally clear that the prototypes were responded to best of all, which may indicate the influence of distance from the category prototype. There is, however, one caveat to be made here. If the feedback given on all test trials provided the opportunity for some learning, then because each prototype was repeated eight times, whereas each exemplar only occurred once, there may have been more opportunities to learn to classify the prototypes correctly than the exemplars (far or near), and this may have been sufficient to explain participants' superior performance on them. In Experiment 2 we modified the procedure on test trials to rule out this possibility.

Experiment 2

Method

To provide a fairer comparison between performance on prototype and on the far exemplar trials, we tested participants on only a single (randomly chosen) far exemplar. This was one change in procedure for Experiment 2. We also tested participants with or without feedback on these two kinds of test trial to assess whether any differences in performance might be dependent on the opportunity to learn. However, to ensure that participants did not forget which class of stimulus belonged to which category, feedback was maintained throughout the test on all trials with close exemplars. We also made the training discrimination rather harder than in Experiment 1. One observation from studies on the peak shift is that although some neighboring stimuli are responded to more rapidly than S+, the effect disappears as the stimuli become further removed from S+. The range of stimuli to which participants respond more rapidly in fact depends on the separation between S+ and S− in training. In Hanson's (1959) study, for example, when pigeons discriminated between 550 nm (S+) and 590 nm (S−), they responded more rapidly only to 540 nm, but when trained to discriminate between 550 nm and 555 nm, they also responded more rapidly to stimuli of 530 and 520 nm than to S+. It is conceivable that, in Experiment 1, the prototype stimuli were at the optimal distance from the training exemplars to yield the peak shift, whereas the far exemplars were too far away from the other category. In Experiment 2, therefore, we trained participants on two different discriminations, both with somewhat more overlap of common elements than the stimuli used in Experiment 1. One final change consequent on this was that we introduced participants to the experiment by training them initially on a very easy discrimination with low overlap between the exemplars; we anticipated that this would minimize subsequent drop-out rates.

Participants and apparatus. There were 29 participants drawn from the same population as in Experiment 1, but because the experiment was run over two sessions they were paid for their participation. The apparatus was exactly the same as before.

Procedure. Each of the two sessions of training began with training on a very easy discrimination by using very low-overlap prototypes (with different stimuli on each day), followed by a second discrimination plus test trials. The prototypes and training exemplars for this preliminary problem are described in Table 3. All participants received a minimum of 20 and a maximum of 80 trials on each of these preliminary problems, stopping after 20 trials as soon as their total percentage correct was significantly ($p \leq .05$) above chance. All participants reached this criterion on both problems within the 80-trial limit. Following this preliminary problem, the session continued with a problem involving either a medium or high overlap between the category prototypes. Half of the participants learned the medium-overlap problem in Session 1 and the high-overlap problem in Session 2, and half learned in the reverse order. The overlap between prototypes and training exemplars for each problem is shown in Table 3, where it can be seen that the medium-overlap problem contained more overlap than that used in Experiment 1. Only six squares per row were changed to generate the second prototype from the first in this experiment, rather than the seven squares per row used for the low-overlap problem in Experiment 1. The high-overlap problem had even greater overlap between the prototypes, with only five squares per row changed in Prototype 1 to construct Prototype 2. Training continued for 60 trials, followed by two blocks of 40 test trials. Within each test block there were 24 retraining trials with more close exemplars, eight prototype trials, and eight trials on two randomly chosen far exemplars (one from each category). Half of each type of trial was with stimuli from each category. Within each block of five trials, there was one prototype, one far exemplar, and three close exemplars.

The data from 5 participants who scored 50% or less correct during test on their close exemplar retraining trials in either session were discarded, leaving 24 participants in the experiment, 12 of whom received feedback on all test trials, whereas the remaining 12 received feedback only on the close exemplar retraining trials. All other procedural details not specifically mentioned were exactly the same as in Experiment 1.

Table 3
Algorithm for Generating the Prototypes and Exemplars for the Two Categories Used in Experiment 2

Characteristic	Prototype 1	Prototype 2
	Design	
Very low overlap	Random	Ten squares per row of Prototype 1 changed
Medium overlap	Random	Six squares per row of Prototype 1 changed
High overlap	Random	Five squares per row of Prototype 1 changed
	Exemplar[a]	
Close	Change 25 of the squares that differ between Prototypes 1 and 2	Change 25 of the squares that differ between Prototypes 1 and 2
Far	Change 25 of the squares that Prototypes 1 and 2 share in common	Change 25 of the squares that Prototypes 1 and 2 share in common

[a]Applies to exemplars at all overlaps.

Results

As in Experiment 1, two measures of performance on test trials were analyzed: percentage correct responses and reaction time. The results are shown in Figures 2 and 3, averaged across the two groups—those that received feedback on all trials and those that received feedback only on retraining trials with close exemplars. Both groups were, of course, treated exactly alike (and behaved similarly) on the close exemplar trials. However, an ANOVA of their performance on prototype and far exemplar trials revealed no difference between the two groups and no interaction between groups and any other variable ($F_{max} = 2.35, p > .10, MSE = 137$).

Inspection of these figures suggests that participants responded more accurately and more rapidly to prototypes than to far exemplars and more accurately and somewhat more rapidly to far than to close exemplars. These differences do not appear to be affected by the degree of overlap between the prototypes on the initial discrimination—indeed, this factor had only a small effect on performance.

Because the experiment was designed to provide a comparison between prototype and far exemplar trials, our initial analysis was confined to these two classes of trial. An ANOVA confirmed that participants responded more accurately and more rapidly to the prototypes, $F(1, 22) = 20.00$ and $16.74, p < .01, MSE = 137$ and 22,308, respectively, but that there was no effect of prototype overlap nor any interaction involving this variable on either measure ($F_{max} = 1.67, p > .10, MSE = 343$). Further analyses, which included performance on close exemplar trials as well, still revealed no effect other than type of stimulus, and Newman–Keuls tests revealed that participants responded significantly more accurately, but not more rapidly, to far than to close exemplars. Although the difference in overlap between the medium and high conditions was not

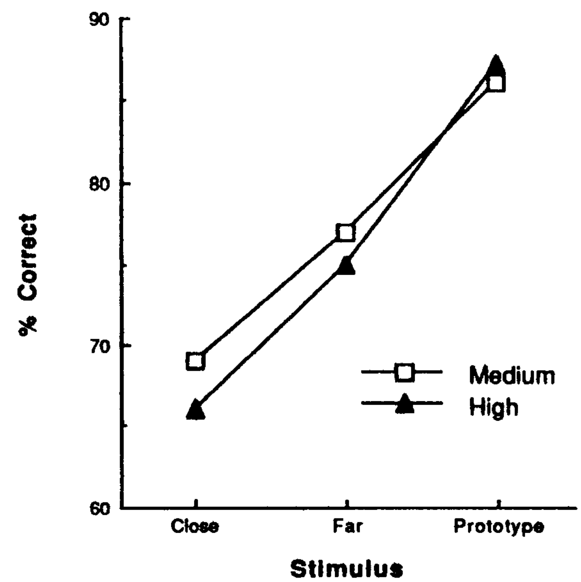

Figure 2. The results for Experiment 2 on the percentage correct measure.

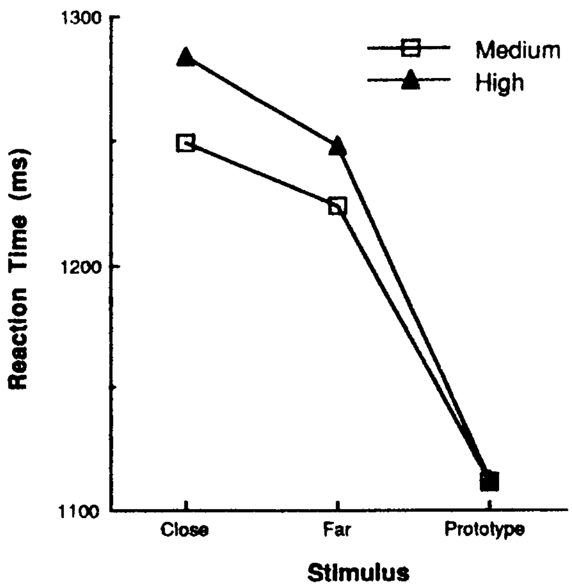

Figure 3. The results for Experiment 2 on the reaction time measure.

great, we had still expected to see some effect of this manipulation. It is true that, except on the two prototypes, performance was better on the medium- than on the high-overlap stimuli, but the differences were very small and far short of significance. We also examined performance during initial training on the two discriminations, but here again there was no effect of this manipulation (Fs < 1).

Discussion

Taken together, the results of Experiments 1 and 2 establish that, under the conditions of our experiments, prototypes are easier to classify as members of a category than far exemplars (Experiment 2), and far exemplars are easier to classify than close ones (Experiment 1). The finding that far exemplars, which differ from close exemplars in their distance from the prototype of the other category (but not from the prototype defining their own category), are easier to classify than close exemplars is important. It throws into sharp relief the role that an exemplar's relationship to the alternative category can play in making decisions on category membership.

Prototypes are defined entirely in terms of the distribution of exemplars of their own category. Hence, if we wish to speak of a "prototype effect," whereby prototypes are more easily classified as members of a category than other exemplars, it behoves us to ensure that such an effect is not dependent on the distribution of exemplars in the alternative category. Given our results, this could have contributed to the prototype effects observed in experiments such as that of Medin and Schaffer (1978).

The finding in Experiment 2 that prototypes are classified more easily than the far exemplars, however, suggests that we have a "true" prototype effect, which is not attributable to the fact that the prototype for Category 1 shares fewer features in common with the training exemplars of Category 2 than do the far exemplars of Category 1. If that is true, we have a prototype effect that cannot be explained as an instance of a peak shift. However, because Hanson's (1959) results establish that the range of stimuli that are responded to more accurately than S+ depends on the distance between S+ and S−, this conclusion requires the demonstration that prototypes are responded to more accurately than far exemplars across a range of overlaps between the two categories. Although in Experiment 2 we attempted to provide such a demonstration, it was hardly an unqualified success because the manipulation of overlap between the two categories produced no significant effects on performance at all. This issue is taken up in Experiment 3.

Whether or not this is the first study to show a true prototype effect, it is the first that has systematically assessed the interaction between prototype and peak-shift effects. This also brings us to the question of possible explanations for both effects. At this point we consider one general class of theory, exemplar theories of the type espoused by Medin and Schaffer (1978) and Nosofsky (1986, 1987, 1989, and especially, 1991, for some relevant discussion of the issues considered in this article) by way of an introduction to Experiment 3. What follows is intended to provide some general characterization of this type of theory and how it might apply to the experiments considered in this article, rather than being an application of any specific theory. Our intention is to raise issues that the proponents of any particular theory might address at a later date.

An exemplar theory requires that each exemplar be stored in memory when encountered along with some designation of category membership. When a new probe exemplar is presented, its similarity to each of the stored exemplars is computed and weighted by factors such as salience, frequency of presentation, and recency of presentation as well. These similarities are summed separately for exemplars from each category, giving two scores denoting overall similarity to either category. The similarity score for one category (S1) divided by the sum of the two overall scores (S1 + S2) is taken as the probability, or more generally an index of the probability, of deciding that the probe exemplar is a member of that category. Nosofsky (1986, 1987, 1989, 1991) has developed a detailed theory that is based on this approach. He uses multidimensional scaling on identification data to determine the "psychological distance" (d) between the stimuli used in his experiments and then converts distance into similarity by using a function of the form similarity, $S = \exp(-d^p)$, where p can be 1 or 2, depending on the confusability of the stimuli used in the experiment. Nosofsky takes p to be 1 when the stimuli are readily discriminable and $p = 2$ when they are highly confusable (see, e.g., Nosofsky, 1989). In the absence of a multidimensional scaling solution for the stimuli used in this experiment, we can only speculate on the application of this theory to our data. In principle, Nosofsky's approach could generate a true prototype effect and a peak shift; the former by virtue of the prototype's greater similarity to the set of training exemplars from its category than any other exemplar, the latter by appealing to the trade-off between an exemplar's similarity to the training set drawn from its own category and the training set drawn from the other. It would simply require that there be the appropriate distances in the psychological space to be converted into similarities by means of one function or the other. However, if these similarities were to reflect the overlap (in terms of number of squares shared) between the test stimuli in a relatively straightforward way, then the method of generation for the stimuli used in these experiments would lead to problems for the theory. Inserting similarities derived from measures of overlap into the decision rule given above does not rank order the stimuli correctly, for the theory then predicts that performance on the far exemplars should be at least as good as on the prototypes.

To see this, we assumed that training involves n_1 close exemplars from Category 1 and n_2 from the other category. The probability that the prototype of Category 1 will be classified as such is given by

$$P(\text{Category } 1/\text{Prototype } 1) = n_1 s_1/(n_1 s_1 + n_2 s_2) \quad (1)$$

on the assumption that s_1 is the similarity of the prototype to any close exemplar drawn from its category and that s_2 is its similarity to close exemplars from the other category.[1] Similar formulas give the probability of correct classification for close and far exemplars. To derive predictions from Equation 1, it is necessary to have a measure of similarity. One simple possibility used by Pearce (1987) to construct a model of conditioning and generalization is that the proportion of features shared by two stimuli serves as an index of similarity. This measure varies between (0) *entirely different, no shared features* and (1) *same, all features shared* as required and captures the intuitively plausible idea that similar things possess many of the same characteristics. If we apply this method of computing similarity to the simple case in which a feature is a black or white square, then we can predict which of prototypes or far exemplars will be easier to classify correctly. The far exemplars in Experiment 1 (for example) shared 206 squares with close exemplars of their own category and 144 with those from the other category. The equivalent figures for prototypes were 231 and 169, and as there were 256 squares in total, we can compute proportions and substitute them into Equation 1. The result is that, contrary to the results of Experiments 1 and 2, the far exemplars (P = .589) are predicted to fare better than the prototypes (P = .578).

Another possibility is that the proportion of squares shared by two checkerboards is better reflected by the psychological distance between the two stimuli rather than a direct measure of their similarity. As a rough example, consider the case in which d is taken to be given by the number of squares on which two checkerboards differ. Each far exemplar differs by 50 squares from close exemplars of its own category and from the other category's close exemplars by 25 squares more than its prototype does. The number of squares by which a close exemplar differs from its prototype is 25, and it differs from the other prototype by 25 squares less than the amount by which the two prototypes differ. For example, in Experiment 1 the prototypes differed by 112 squares, and so a close exemplar's distance from the other category prototype would be taken as $112 - 25 = 87$ squares, and a far exemplar's distance from a close exemplar drawn from the other category would be $112 - 25 + 25 = 112$ squares. Conversion of these distances into similarities with p taken as 1 or 2 and substituting them into Equation 1 predicts that prototypes and far exemplars should be equally easily classified with $p = 1$, but that far exemplars should be more easily classified than prototypes if $p = 2$. For example, we have

$$P(\text{Category } 1/\text{Prototype } 1)$$
$$= n_1 \exp(-25)/\{n_1 \exp(-25) + n_2 \exp[-(112 - 25)]\} \quad (2)$$

on the assumption that $p = 1$ in the similarity equation. The relevant probability for a far exemplar from Category 1 is

$$P(\text{Category } 1/\text{Far } 1) = n_1 \exp[-(25 + 25)]/$$
$$\{n_1 \exp[-(25 + 25)] + n_2 \exp[-(112 - 25 + 25)]\}. \quad (3)$$

Use of the fact that $\exp[-(25 + 25)]$ can be written as $\exp(-25) \cdot \exp(-25)$ gives a factor of $\exp(-25)$ throughout Equation 3, which cancels to give P(Category 1/Far 1) = P(Category 1/Prototype 1) independent of category (i.e., prototype) separation. If $p = 2$ instead, then the algebra is a little more complex, but the result is equally unambiguous. P(Category 1/Far 1) is now greater than P(Category 1/Prototype 1), and the far exemplar should be better classified than the prototype.

This raises the question of whether exemplar theories are, in principle, capable of accommodating our results. It would obviously be premature to argue that they cannot. As we have seen, Nosofsky (1986, 1987, 1989, 1991), following Shepard (1987), used psychological distance, on the basis of the results of multidimensional scaling, to measure similarity, and it may be that similarities derived from distances obtained in this way for our stimuli would fit our data. For this to be true, it would be necessary for these distances not to correspond in any simple fashion to the number of squares on which two checkerboards differ and to be such as to generate similarities that departed significantly from those given by the overlap measure. This is perhaps not entirely unlikely, however, as the equation of feature with black or white square is, surely, far too simplistic, and similarity may be better described as some function of the overlap rather than the raw score used here. Given this, it seemed worthwhile including an identification

[1] Taking the similarity between prototype and any close exemplar drawn from a given category to be constant is not unreasonable if similarity is determined by the proportion of features shared by two stimuli. Taking the simplest possible example, in which a feature is taken to be a black or white square, then as each close exemplar shares (taking values from Experiment 1) 231 squares with its prototype and 169 with the other, s_1 and s_2 would be constant. More generally, as the close exemplars are all constructed in exactly the same way, by changing 25 squares unique to the prototype, then we can expect the similarity between prototype and any close exemplar from a given category to be constant if the perceived psychological similarity directly reflects the physical make up of the stimulus.

component in our final experiment, enabling us to derive a multidimensional scaling solution to see whether our overlap measure captured the similarities between the stimuli as perceived by our participants.

Experiment 3

Method

This time all of the test stimuli were placed on an equal footing by using two prototypes: two far exemplars (one from each category as in Experiment 2) and two close exemplars as test stimuli. Feedback was given to all stimuli, as Experiment 2 indicated that this does not unduly influence the results obtained with our procedures. The two close exemplars were drawn from the training set used in the earlier training phase of each condition. They were the exemplars that had occurred just before criterion was reached, and again there was one from each category. This allows comparison between performance on all three types of stimuli, placing maximal constraints on any theory's ability to account for the data.

We also increased the power of the overlap manipulation by changing the medium-overlap version of the task used in Experiment 2 to a low-overlap condition just like the stimuli in Experiment 1. The lack of any significant influence of the overlap variable in Experiment 2 may well have been due to the small difference in overlap of the prototypes defining the categories in the medium and high conditions. By doubling that difference, we hoped that an effect of difficulty could be detected.

Participants and apparatus. There were 42 participants drawn from the same population as before. The apparatus was unchanged.

Procedure. As in Experiment 2, training on the low and the high conditions was preceded by preliminary training on an easy problem with very low overlap designed to introduce participants to the task. Participants were then required to complete a minimum of 30 trials and a maximum of 100 during training on both the low- and high-overlap conditions. They could terminate training before the 100 trials were up by meeting a criterion, which was performance better than chance at $p \leq 0.1$ and $p \leq 0.05$ for the low- and high-overlap conditions, respectively. These criteria were chosen in the hope of ensuring that performance on test would be roughly equal in the two conditions. The test phase involved two blocks of 24 trials, with each stimulus occurring four times per block and once every six trials—the order of stimulus presentation within each group of six trials being random. Otherwise the procedures in the categorization phase of this experiment were exactly as in Experiment 2. Analyses were carried out both for all of the participants, regardless of their performance on test, and after dropping those participants who scored 50% or less correct during test on the close exemplars. Because this did not materially affect the results, the data for all 42 participants are reported.

Multidimensional scaling was carried out in a separate session for 16 of the participants who were able to return for this phase of the experiment, after the categorization phase was completed. An identification task was used to generate the raw data. Participants were shown six stimuli: two prototypes, two close exemplars, and two far exemplars; for half of the participants they were from the low-overlap condition, for the other half they were from the high-overlap condition. These stimuli were novel for these participants, although drawn from the pool of stimuli used during the main experiment. Each stimulus was matched to a number (1–6), and participants were told that they would have to press the key of that number when that stimulus appeared on the screen. There was no time pressure, and feedback was given (either correct or the correct number). Each block of trials cycled the participant through the six stimuli six times in random order. There were four blocks, with the first counting as practice. All of the participants were then shown another six stimuli drawn from the opposite condition, and the whole procedure repeated. The confusion matrices from this identification phase were then analysed separately for the low- and high-overlap stimulus sets.

Results

Figures 4 and 5 give the percentage correct and reaction time data for the categorization phase of this experiment. Accuracy of performance was higher on prototypes than on far exemplars, which were in turn classified more accurately than close exemplars, even though the close exemplars were recently experienced members of the original training set. For the percentage correct data, there was a significant effect of stimulus type, $F(2, 82) = 28.8, p < .001, MSE = 280$, and an interaction between stimulus type and difficulty, $F(2, 82) = 3.22, p < .05, MSE = 260$. All other Fs were nonsignificant. Further Newman–Keuls analyses confirmed that performance on prototypes was significantly better than that on far exemplars, which were categorized significantly more accurately than close exemplars ($p < .05$ at least). Inspection of the figures suggests that the significant interaction between stimulus type and difficulty was due to the poorer performance on close and far exemplars in the high-overlap condition, contrasted with roughly equal performance on the prototypes. Newman–Keuls tests established a significant difference between the two conditions on the close exemplars ($p < .05$), but no difference on the far exemplars.

The reaction time data do not significantly contradict the accuracy data and emphasize the effect of the difference in overlap between the two prototypes. The only significant effect was for overlap, $F(1, 41) = 7.86, p < .01, MSE = 484{,}379$, with the high condition taking longer than the low. Overall, there

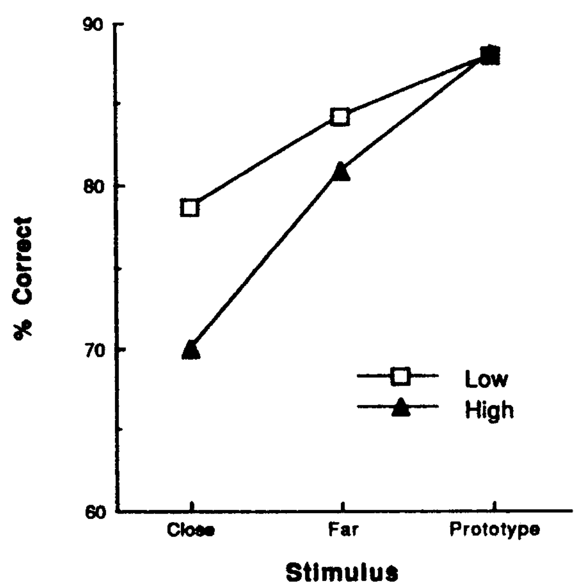

Figure 4. The results for Experiment 3 on the percentage correct measure.

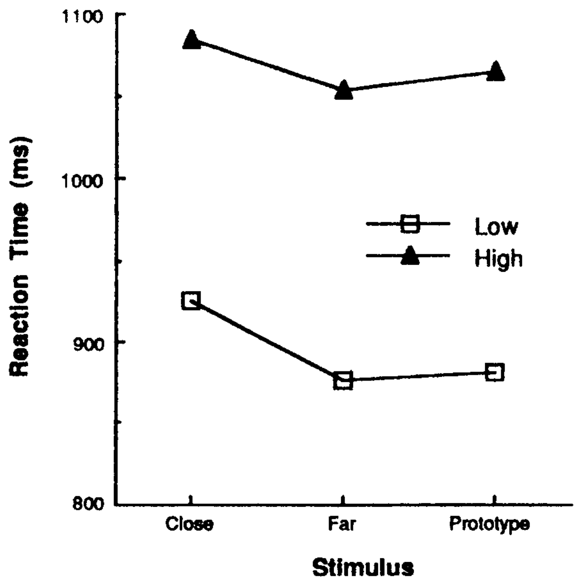

Figure 5. The results for Experiment 3 on the reaction time measure.

can be little doubt that prototypes were categorized more easily than far exemplars, which were in turn categorized more easily than close exemplars, thus confirming the results of Experiments 1 and 2. The manipulation of overlap between prototypes defining the categories to be discriminated also affected performance here, increasing reaction times to all test stimuli and decreasing accuracy on close and far exemplars, even though the prototypes were well matched for accuracy.

Multidimensional Results

The confusion matrices (shown in Table 4) obtained from the identification phase were scaled by using SYSTAT to give solutions in one and two dimensions. The two-dimensional solutions accounted for little extra variance, and hence the one-dimensional solutions were preferred. The actual percentage of the variance accounted for by these solutions was 90% for the low stimuli, and 99% for the high stimuli, and the results of the scaling are shown in Figure 6. These were readily interpretable as corresponding to the overlap measure (proportion of squares shared) in that stimuli sharing less squares are generally further apart. The six stimuli used in the multidimensional scaling phase at each level of category separation can be rank ordered in terms of distance on the basis of their overlap, allowing correlation with the rank ordering generated by multidimensional scaling. This gives $r_s = .94$ and $r_s = .83$ (both $ps < .05$) for low- and high-overlap stimuli, respectively. Considering the relatively small number of participants available for this phase of the experiment, these correlations provide good support for the contention that the stimulus overlap measure reflects the similarity of the stimuli as perceived by the participants.

There are some caveats to the multidimensional scaling analysis performed here. We scaled a small number of stimuli representative of those used on test in an effort to gather evidence for or against the notion that perceived similarity would be a function of stimulus overlap. We expect that attempts to scale the whole set of training stimuli used in this experiment (if it were possible) would result in a complex multidimensional solution that would probably be uninterpretable. It is hard to see what simple solution could allow for the fact that all close exemplars are equidistant from their prototype and closer to the other prototype than their own prototype is, but at the same time the prototype is the central tendency of these close exemplars. Indeed, it may be that the structure of the stimulus set used in these experiments cannot be captured by either Euclidean or city-block metrics without positing an unreasonably large number of dimensions. Attempting to find low-dimensional solutions that reflect the constraints of the overlap measure illustrates this problem. In the case of the far exemplars, a solution that is based on a city-block metric is possible. It can be imagined that each prototype is surrounded at some distance by far exemplars scattered around a circle in a plane orthogonal to the line joining the two prototypes. Each far exemplar is the same distance (say d) from its prototype, the central tendency of these exemplars is the prototype, and each far exemplar is the same distance as its prototype from the other prototype plus d. In the case of the close exemplars, however, the problem is that they are not only all equidistant from their own prototype but are also closer to the other prototype than their own prototype is.

How can this be represented? If all of the close exemplars are closer to the other prototype than their own prototype is, averaging them will produce a point that does not fall on their prototype but will be shifted toward the other category prototype. The only escape would seem to lie in either adopting a metric that allows the average distance of a set of

Table 4
Confusion Matrices for Experiment 3

Prototype or exemplar	F1	P1	C1	C2	P2	F2
			Low overlap			
F1′	.979	.006	.006	.000	.006	.000
P1′	.011	.944	.024	.017	.027	.003
C1′	.004	.022	.944	.017	.003	.003
C2′	.000	.018	.017	.937	.013	.017
P2′	.003	.017	.003	.015	.927	.027
F2′	.000	.002	.003	.013	.024	.947
			High overlap			
F1′	.916	.031	.020	.017	.006	.003
P1′	.048	.902	.017	.010	.006	.003
C1′	.010	.062	.937	.031	.003	.003
C2′	.013	.003	.020	.909	.010	.000
P2′	.006	.000	.000	.020	.940	.017
F2′	.003	.000	.003	.010	.031	.972

Note. F, P, and C stand for far exemplar, prototype, and close exemplar, respectively. A 1 or 2 designates the category that the stimulus belonged to, whereas ′ indicates that this was the response given by the participant. For example, the entry for high-overlap P1, C2′ is .003, meaning that the prototype for Category 1 was identified as the close exemplar from Category 2 on 0.3% of the trials averaged across participants.

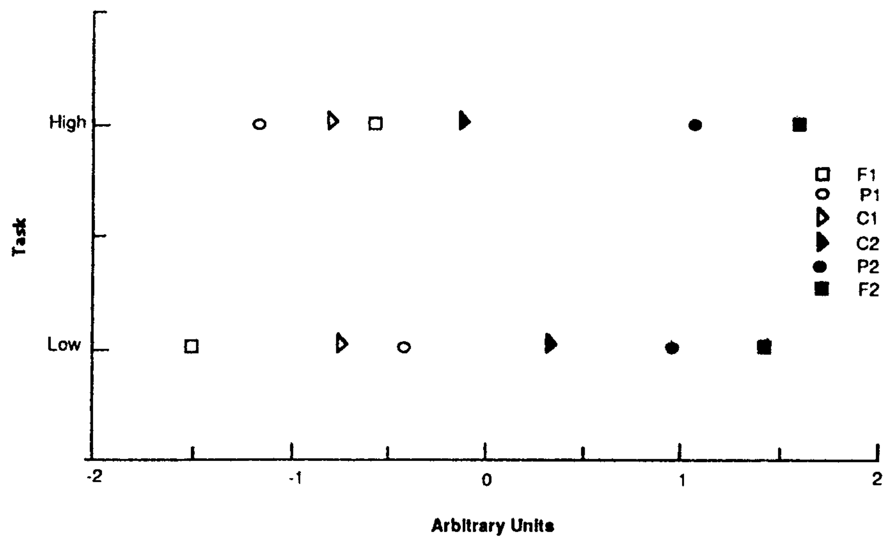

Figure 6. The results of multidimensional scaling for the test stimuli used in Experiment 3. F = far exemplar; P = prototype; C = close exemplar.

points all a certain distance (d) from some reference point to be greater than d or by using a high-dimensional solution that encodes the structure of each checkerboard stimulus in a fairly direct fashion. An example of the latter would be a 256-dimension space in which each dimension codes whether a particular square is black or white.

In any event, our central claim is that the strong correlation between stimulus overlap and perceived similarity needs to be explained. Our explanation is simple, that one drives the other and that it follows from this that exemplar theories of the type considered here will have difficulty in fitting the data gathered on test. Really though?

General Discussion

The results of Experiments 1 to 3 can be summarized as showing that, by using our stimuli and procedures, prototypes were categorized better than far exemplars, and far exemplars were categorized better than close exemplars. Can this order ever be reversed? We have attempted to see whether such a shift in the rank ordering of prototypes and far exemplars could be produced by increasing the overlap between the two prototypes and reducing the distortion used to create far and close exemplars. However, either we have observed no reversal or the problem has been too difficult for our participants to solve. This emphasizes the reliability of the results reported in this article and suggests that with these stimuli they are general rather than exceptional.

The advantage enjoyed by the far exemplars over the close is a clear example of a peak shift because both types of stimulus were the same distance from the prototype. However, the finding that prototypes were better classified as members of their category than far exemplars drawn from that category indicates that there was more than just a peak shift at work in these experiments. The only way to resist this conclusion would be to argue that the peak of the postdiscrimination gradient just happened to fall at the location of the prototypes and that the gradient then declined by the time the far exemplars were reached. Our manipulation of the separation of the two categories across the three experiments was designed to rule out this possibility because experiments on peak shift have established that the precise location of the peak of the postdiscrimination gradient varies as a function of the separation between S+ and S−. In fact, the interaction observed in Experiment 3 was in precisely the opposite direction from that expected on such an analysis. In Hanson's (1959) pigeon experiment, a decrease in the separation between S+ and S− was accompanied by a further shift in the peak of the gradient away from S+. Translated into the present case, this implies that an increase in overlap between the two categories in Experiment 3 should have improved performance on the far exemplars relative to the prototypes. In fact, it appeared to increase the superiority of the prototypes.

This true prototype effect presumably occurred because the prototype was the central tendency of the training exemplars; even though all of the training exemplars were "nearer" the other category than the prototype, the prototype was the average of all those exemplars. This seemingly paradoxical state of affairs simply reflects the method of generation of the close exemplars—by adding noise to the prototype. The finding that prototypes were classified more accurately than far exemplars gives us one kind of prototype effect: an advantage for prototypes relative to new exemplars. It also implies that the prototypes' advantage over close exemplars in these experiments could not have been due to a peak shift alone. If this is true, then these experiments provide evidence for the most powerful type of prototype effect: an advantage for the prototype over old exemplars. The argument rests on the assumption that given the range of separation between the categories we allowed, peak shift alone would have resulted in better performance on far exemplars than on prototypes in at least one of the discriminations used in Experiments 1–3. Hence, an upper estimate of the contribution of a peak shift to the difference between prototypes and close exemplars is given by the difference between far and close exemplars. Because the prototypes were classified significantly better than far exemplars at all levels of category separation, it follows that the advantage enjoyed by prototypes over the close exemplars was significantly greater than would be expected on the basis of peak shift alone.[2]

One implication of these results, then, is that not all prototype effects reported hitherto are equal. At least some are probably due to an unknown combination of two factors: a central tendency effect and a peak shift. Categorization of a test stimulus is determined not only by the relationship of that stimulus to its own category but also by its relationship to any alternative category. A true prototype effect is that which is left when the peak-shift effect is subtracted out; that is, it should be independent of the distribution of the alternative category and depend only on the distribution of its own training exemplars. Posner and Keele's (1968) demonstration of an advantage for prototypes over new exemplars meets this criterion because their exemplars were random distortions of the prototype and so should not, on average, have benefited from a peak-shift effect. Despite an extensive survey of the literature, however, we have not been able to find any demonstration that prototypes are classified significantly more easily than old exemplars in which one can be confident that no peak-shift effect has contributed.

What remains to be considered are the implications of these results for theories of categorization. As we have seen, it seems that an exemplar theory of the type considered here is constrained to predict that far exemplars will always be categorized at least as well as prototypes; in other words, that the proximity of a stimulus from one category to the set of training exemplars from the other category is at least as important a factor as its proximity to the central tendency of its own category. This may seem surprising for a class of theory that is, in part, precisely designed to explain prototype effects. However, although we have not fully considered the consequences of allowing changes in attention to some features in preference to others, we believe that other versions of exemplar theory will have equal problems with our results. Our reasoning is that this type of theory is predicated on the notion that stimuli can be represented in a psychological space, with similar stimuli closer together in the space. The problem lies in capturing our stimulus structure in such a space, in particular, in allowing close exemplars all to be nearer the prototype of the other category than their own prototype is, although still having that prototype as their central tendency.

As an alternative, consider an elemental approach in the spirit of McClelland and Rumelhart (1985), McLaren, Kaye, and Mackintosh (1989), and Shanks (1991). Assume that any category exemplar has a conglomeration of features represented as a set of active elements. Assume also that on each trial these active elements become associated with category membership. We can start by using the crude approximation of taking a feature to be whether a square in a given position is black or white. On this scheme, prototypes would share x elements in common and there would be y $(= 256 - x)$ elements unique to each prototype. On each training trial, $y - 25$ unique elements from the correct prototype and 25 from the incorrect can potentially be associated with category membership because 25 squares unique to a prototype are changed to generate the close exemplars used in training. On average, then, the y unique elements will come to predict category membership. The x common elements will be equally associated with both categories and so will not be predictive of category membership. On this simple model, close exemplars, which have 25 unique elements exchanged for 25 elements from the other category, clearly have less net associative strength for correct category membership than do prototypes. A close exemplar will have $y - 25$ elements associated with the correct category and 25 with the incorrect one, giving $y - 50$ net in favor of the correct decision. Prototypes will have y in favor of the correct category and none against. If decisions are made on the basis of this strength, close exemplars would be classified less readily than prototypes. Far exemplars, however, would be classified as easily as prototypes. They possess the same set of unique elements in their representation, having exchanged some common elements equally associated with both categories for others that possess no associations to either.

Is there any way of saving this approach? One possibility is to reject, as too simplistic, the assumption that individual black and white squares are the only features. Configurations of squares can be expected to play an important role and to be represented as well. In particular, configurations involving a

[2] If this is not immediately clear, a rough attempt to quantify this claim may help. Consider what pattern of results would be expected if peak shift were the only factor operating. Decreasing category separation has clearly brought the far exemplars close enough to the opposing category to suffer from its effects, as witnessed by the drop-off in performance. Given this, the prototypes should deteriorate in performance by more than the amount observed for the far exemplars. We can take the decline in performance of the far exemplars as a lower bound on the drop-off to be expected (on a peak-shift account) in performance on the prototypes, and a reasonable estimate of the drop-off to be expected for the prototypes is obtained by interpolation between that observed for far and close exemplars. This is justified by the fact that far and close exemplars were equidistant either side of the prototypes on the overlap measure and roughly equidistant on our scaling solution. Taking this as our estimate of the expected decline in performance for the prototypes (on a pure peak-shift account), it follows that the observed difference between prototype performance and our expectations is an estimate of that component of prototype performance not accounted for by peak shift and that this component is the excess over close exemplar performance that cannot be due to peak shift. Taking a mean square error term based on all the data, this component has an associated $F(1, 41)$ of 4.09, $p < .05$, $MSE = 260$. Thus our calculations support the claim that there is a significant advantage for prototypes over close exemplars. The estimate of the peak-shift component in the advantage enjoyed by the prototypes over the close exemplars is only slightly (1%) greater than the far–close difference for the high-overlap condition in this example.

combination of squares unique to a given prototype and squares common to both prototypes could actually serve to discriminate between categories; hence elements representing such combinations will gain some associative strength. Some of these configural features are lost when squares common to both prototypes are changed to generate a far exemplar, hence far exemplars can be expected to be classified less easily than prototypes. Furthermore, close exemplars can still be expected to be more difficult than far exemplars, as they will suffer from a loss of individual squares that are associated with category membership and a loss of configural features deriving from combinations of unique squares and combinations of unique and common squares, both of which are predictive of category membership.

To see this, consider the case in which pairwise configurations of squares are taken to be important as well as individual squares. When 25 squares of a prototype are changed to create a far exemplar, then 25×112 pairwise configurations involving one square unique to the prototype are lost (assuming the prototype has 112 unique squares as in the low-overlap case). The elements representing these configurations would differentially be associated with category membership, and so their loss should result in a deterioration in performance. The argument generalizes to higher order configurations without difficulty.

With these assumptions, then, the elemental model is able to predict both that prototypes will be classified more easily than far exemplars and that far exemplars will be easier than close exemplars. The former difference hinges on a loss of configural cues, the latter on changes in features relating to individual squares as well. Given adequate training with a sufficient number and distribution of exemplars to define the category prototype, it is difficult to see how this model could ever predict a reversal of prototypes and exemplars of either kind in this type of experiment.

This raises the question of whether exemplar theories would not also benefit from assuming that stimulus similarity is determined by the proportion of shared configurations. We have carried out an analysis of similarities computed in this way for various orders of configural cue, but insertion of these similarities into Equation 1 still leads to the prediction that far exemplars will be better classified than prototypes. The prediction remains unaffected even if we use all possible configurations simultaneously; nor is it changed by using a square or square root transform, nor even by multiplying similarities at each level of configuration together to get an overall similarity measure (as suggested in Medin & Schaffer, 1978). It would seem that exemplar theories are not able to benefit as readily as the elemental theory espoused in this article from the inclusion of configural cues in the analysis.

Although a simplistic application of an elemental theory seemed to predict that far exemplars would be categorized as well as prototypes, it was readily rescued by the wholly plausible assumption that configurations of black and white squares provided discriminable features that served to differentiate the training exemplars. Far exemplars would have preserved fewer of these configurations than the prototypes, and this loss should be sufficient to explain why they were categorized less accurately at all levels of separation between the two categories. This account could also allow one to explain the interaction observed in Experiment 3 between level of separation and relative performance on the three types of test stimuli. Although the prototypes were classified equally well at both levels of separation, performance on far exemplars was somewhat worse in the high-overlap condition, and this difference was even more marked for close exemplars. Now because the number of elements differentiating the two categories will be less the greater the overlap between the two categories, it follows from an elemental theory that each of these elements will be carrying more associative strength after training on the high-overlap task. Because exemplars were generated by changing the same number of squares (25) from the prototype in both conditions, the loss of associative strength would have been greater in the high-overlap condition. The close exemplars would have lost associative strength from cues corresponding to individual squares and configurations of them. The far exemplars would have been less affected, however, because only elements representing configural features predictive of category membership were lost.

The elemental account of categorization is both simple and seems successful in predicting the observed pattern of results. A more detailed appraisal of its strengths is now needed, which in turn demands that a suitably detailed and explicit model instantiating this approach be constructed.

References

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

Hanson, H. M. (1959). Effects of discrimination training on stimulus generalization. *Journal of Experimental Psychology, 58*, 321–334.

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

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

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

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

McLaren, I. P. L., Kaye, H., & Mackintosh, N. J. (1989). An associative theory of the representation of stimuli: Applications to perceptual learning and latent inhibition. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology* (pp. 102–130). London: Oxford University Press.

Medin, D. L., Altom, M. W., & Murphy, T. D. (1984). Given versus induced category representations: Use of prototype and exemplar information in classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 333–352.

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

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

Nosofsky, R. M. (1987). Attention, and learning processes in the identification and categorization of integral stimuli. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, 13, 87–108.

Nosofsky, R. M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Perception and Psychophysics, 45,* 279–290.

Nosofsky, R. M. (1991). Typicality in logically defined categories: Exemplar-similarity versus rule instantiation. *Memory & Cognition, 19,* 131–150.

Pearce, J. M. (1987). A model of stimulus generalization for Pavlovian conditioning. *Psychological Review, 94,* 61–73.

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

Rilling, M. (1977). Stimulus control and inhibitory processes. In W. K. Honig & J. E. R. Staddon (Eds.), *Handbook of operant behavior* (pp. 432–480). Englewood Cliffs, NJ: Prentice Hall.

Rosch, E. (1978). Principles of categorization. In E. Rosch and B. B. Lloyd (Eds.), *Cognition and categorization.* Hillsdale, NJ: Erlbaum.

Shanks, D. R. (1991). Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 433–443.

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

Wagner, A. R., Logan, F. A., Haberlandt, K., & Price, T. (1968). Stimulus selection in animal discrimination learning. *Journal of Experimental Psychology, 76,* 171–180.

Whittlesea, B. W. A. (1987). Preservation of specific experiences in the representation of general knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 3–17.

Received May 13, 1993
Revision received June 17, 1994
Accepted June 28, 1994 ■