Animal Learning & Behavior
2002, 30 (3), 177-200

Associative learning and elemental representation: II. Generalization and discrimination

I. P. L. McLAREN and N. J. MACKINTOSH
University of Cambridge, Cambridge, England

This paper follows on from an earlier companion paper (McLaren & Mackintosh, 2000), in which we further developed the elemental associative theory put forward in McLaren, Kaye, and Mackintosh (1989). Here, we begin by explicating the idea that stimuli can be represented as patterns of activation distributed across a set of units and that different stimuli activate partially overlapping sets (the degree of overlap being proportional to the similarity of the stimuli). A consequence of this view is that the overall level of activity of some of the units representing a stimulus may be dependent on the nature of the other stimuli present at the same time. This allows an elemental analysis in which provision for the representation of configurations of stimuli is made. A selective review of studies of generalization and discrimination learning, including peak shift, transfer along a continuum, configural discrimination, and summation, suggests that the principles embodied in this class of theory deserve careful consideration and will form part of any successful model of associative learning in humans or animals. There are some phenomena that require an elemental/associative explanation.

In an earlier paper (McLaren & Mackintosh, 2000), we developed and elaborated a theory briefly outlined by McLaren, Kaye, and Mackintosh (1989) and applied the theory to the main phenomena of Pavlovian conditioning, with special emphasis on latent inhibition. In the second half of the paper, we applied the theory to the case of perceptual learning. Our underlying argument was that models that combine associative learning algorithms with elemental representations of stimuli are surprisingly powerful and, perhaps even more surprising given the long history of this approach, contain resources that have still not been fully exploited or analyzed. The detailed application of the model to perceptual learning was intended to illustrate this point. William James (1890) drew attention to the remarkable discriminatory powers of experts in particular sensory domains, such as wine tasters. Experimental analysis has long documented the fact that mere exposure to two or more similar stimuli, in the absence of any instruction or differential reinforcement, can facilitate their subsequent discrimination (see Hall, 1980, for an early review). But theoretical understanding of the mechanisms underlying these perceptual learning effects has lagged a long way behind their empirical demonstration, and they have always seemed particularly resistant to an associative analysis (see Gibson & Gibson, 1955). When exploited to the full, however, we believe that the resources of an associative, elemental approach will turn out to be sufficient to provide a detailed explanation. Such an approach has unquestionably generated a variety of novel predictions.

In the present paper, we will apply our model to the rather more familiar and much more extensively studied phenomena of generalization and discrimination. These are large topics, and our coverage will be far from exhaustive. The main emphasis will be on the elemental aspects of our approach, rather than on a detailed justification of an associative analysis of discrimination learning, although as we shall see, the associative principles invoked by our model play a crucial role in the explanation of certain findings. The main question at issue is how stimuli are represented and whether elemental representations are powerful enough to capture the main phenomena of generalization and discrimination. Thus, it will come as no surprise that much of our discussion, at least of discrimination learning, will focus on the contrast between elemental theories and the type of configural analysis recently and persuasively advocated by John Pearce (1987, 1994).

THE MODEL

Brief Reprise

Details of the formal version were given in our previous paper (McLaren & Mackintosh, 2000) and do not need to be repeated here. We will begin with the general principles underpinning the model and then will develop them further in the context of generalization and discrimination.

Our model assumes an error-correcting learning rule that employs real-time learning, weight decay, and salience modulation. We follow earlier elemental theories of con-

Correspondence concerning this article should be addressed to I. P. L. McLaren, Department of Experimental Psychology, Downing Street, Cambridge CB2 3EB, England (e-mail: iplm2@cus.cam.ac.uk).

—*This article was invited by the editors.*

ditioning, such as stimulus sampling theory (Atkinson & Estes, 1963; Estes, 1959), which conceptualized all stimuli as sets of elements, the elements themselves being simple primitives (corresponding, perhaps, to what would today be termed the microfeatures of a stimulus).

More specifically, we postulate the following.

1. The representation of a stimulus consists of a pattern of graded activation distributed over a set of units corresponding to the elements of the stimulus, rather than there being a one-to-one correspondence between a stimulus and a representational unit.

2. Similar stimuli consist of partially overlapping sets of elements, their degree of similarity being related to the proportion of common elements. Where stimuli can be construed as varying along a continuum or dimension, such as visual wavelength or auditory frequency, different values along the dimension are assumed to consist of a series of overlapping sets of elements (cf. Hull, 1943; Thompson, 1965). In effect, each representational unit is postulated to have a *tuning curve*, responding most strongly to one particular value on the dimension and less strongly to neighboring values. Thus, variation along a stimulus dimension such as wavelength will, for the most part, be represented by different *sets* of units corresponding to different values on the dimension, rather than the activation level of an individual unit being the primary indicator of value on the dimension (Thompson, 1965). Note that many units will be active when any stimulus on that dimension is present; thus, the coding of position on the dimension is in terms of a pattern of activation. When we are dealing with variations in intensity, we assume that increases in intensity are represented not only by increases in the activity of units already active, but also by the recruitment of additional, *neighboring* units. Thus, for both kinds of dimensions, the coding of different values on the dimension is achieved partly by differences in which units are activated and partly by differences in their level of activation.

3. Not all the elements of a given stimulus will actually be sampled during its presentation, and hence, not all of its corresponding units will be activated on a given trial. In addition to the experimenter's nominal conditioned stimulus (CS) or discriminative stimuli, there will be other extraneous stimuli present, which may be sampled in a variable way from trial to trial and which will add noise. In experiments on discrimination learning, the experimenter's discriminative stimuli, S+ and S−, predictive of reinforcement and its absence, respectively, will always be presented in a particular context, and these contextual stimuli, when sampled, will also be associated with the outcome of each trial. As we will see later, this source of noise assumes considerable importance in our theorizing, since the coarsely coded, distributed representations that we assume imply that the units coding a particular stimulus may vary as a function of the other stimuli present at the same time. In addition to external contextual cues, the organism's own internal state generates a constantly varying set of stimuli that will also enter into association with other events, and the experimenter may deliberately introduce incidental stimuli, common to both reinforced and unreinforced trials. All these factors are sources of noise or variability in representation, which usually have to be overcome for successful learning to occur.

Elemental Representations

We now return to the assumptions contained within points 1 and 2. These representational assumptions bear closer examination, since they lie at the very heart of how our model instantiates generalization and discrimination. We start by asking how best to represent stimuli. The answer we have given so far is that stimuli are best represented as sets of elements, with overlap between these sets serving as the basis for generalization. But what, computationally speaking, is the best way of representing stimuli as sets of elements?

Our argument is that a relatively coarse coding is optimal, so that all the stimuli are represented as distributed patterns of activation. To see this, we begin by considering the problem of the most efficient coding scheme for the stimulus space shown in Figure 1, panel A, given that we have four units to span the space. It might be thought that the scheme shown in panel B of the figure would be the best that could be done with four units, since this divides the space into four nonoverlapping regions, allowing us to resolve any stimulus into one of those four regions. But, of course, this scheme allows only four distinctions to be made. Any stimulus that fell into the top left quadrant of the space would activate the same unit. Panel C makes it clear that a better strategy is to use larger *receptive fields* for each unit so that the receptive fields overlap and, thus, allow stimuli to activate combinations of units. This greatly enhances the representational power available, since on this scheme, stimuli can now be represented by 1 of the 16 possible combinations of unit activation: all off (1), any one of them on (4), any pair on (6), any triple on (4), and all four on (1). Only some of these combinations can be illustrated in our figure. The optimal version of this coding scheme would allocate equal areas of the space to each combination, and since each unit is involved in half the possible combinations, this implies that each unit spans half the space.

The remarkable feature of this coarse-coding scheme is that unlike the approach taken in panel B of Figure 1, where, as the number of units increases, the receptive field size decreases and so coding becomes more and more localist in nature, with coarse coding the receptive field size stays at 50% of the space whatever the number of units employed (for n units there are 2^n combinations, of which 2^{n-1} involve any one unit). This means that any arbitrary pair of stimuli will have considerable overlap in terms of the number of units that they activate. Since each will activate a random 50% of the available units,

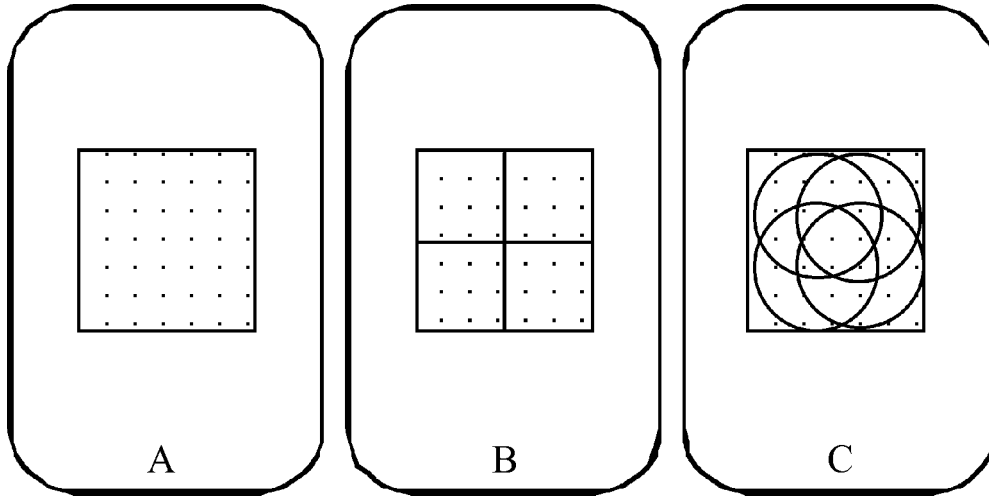

Figure 1. (A) A stimulus space for representation. (B) One method of assigning four units to the space. (C) A more efficient method of making use of the same four units to span the space. See the text for a fuller discussion.

the mean overlap on this scheme is 25%. Two stimulus patterns that have, in some sense, no features in common will still activate overlapping sets of units.

In implementing this scheme, we do not assume a direct, one-to-one relationship between the presence of a stimulus element and the activation of a particular unit. Rather, we assume that each unit is randomly allocated a positive or a negative connection from each element. Statistical fluctuation will then ensure that some units have net positive, and others net negative, input from the elements corresponding to a given stimulus. Indeed, the distribution of input to units will be approximately Gaussian, as is shown in Figure 2A.

Paradoxically, this might appear to pose a problem. If we were to assume a strictly linear relationship between input and activation, with the proviso that units cannot take activations below zero, it would follow that the distribution of inputs shown in Figure 2A would produce the distribution of unit activations shown in Figure 2B: Approximately half have zero activation, with a decreasing number having increasing levels of activation. But we have argued that the representation of any stimulus is an approximately Gaussian pattern of activity distributed across a set of units, as shown in Figure 2C. The resolution to this apparent problem is, of course, to assume that the function relating input to activation is nonlinear, as is shown in Figure 3A.

Whereas for relatively small inputs there is little activation, over an intermediate range of inputs there is an approximately linear increase in activation that then asymptotes. This function will yield an approximately Gaussian shape when applied to our input distribution, because the negatively accelerating activation function at asymptote produces a broader peak than that present in the input distribution. The result is as shown in Figure 3B.

The pattern of activation that we would expect for a novel stimulus is shown in Figure 3B, but what would we expect for a familiar stimulus? Of course, a familiar stimulus would be subject to latent inhibition and so would have reduced input to each representational unit as a consequence of salience modulation. As a result, its distribution of activation over representational units would tend to revert to the underlying roughly exponential pattern of input, as shown in Figure 3C; given the reduced maximum input available, the asymptotic activation possible for a unit would no longer be such a significant factor in determining unit activation. Hence, other things being equal, we would predict that a necessary consequence of familiarization with a stimulus represented in these terms will be a sharpening of its activation profile and, so, a decrease in generalization from it to other stimuli (see the discussion of an experiment from Mackintosh & Little, 1970, below). In other words, familiarization with a stimulus will result in an increase in its discriminability from other stimuli. We have identified a further mechanism underlying perceptual learning effects—in addition to the associative mechanisms outlined in our earlier paper. A variety of studies reviewed by Hall (1991, pp. 58–66) have established that generalization is reduced by extended exposure to the training stimulus.

Stimulus Dimensions

Having considered a single stimulus in isolation, it is now time to consider the question of what constitutes a stimulus dimension. One answer that follows from the scheme advanced here is that a stimulus dimension can be any set of stimuli constructed so that each new neighboring stimulus ($i + 1$) on the dimension is produced by deleting some of the inputs (perceptual elements) of its preceding neighbor (i) and substituting, for the deleted

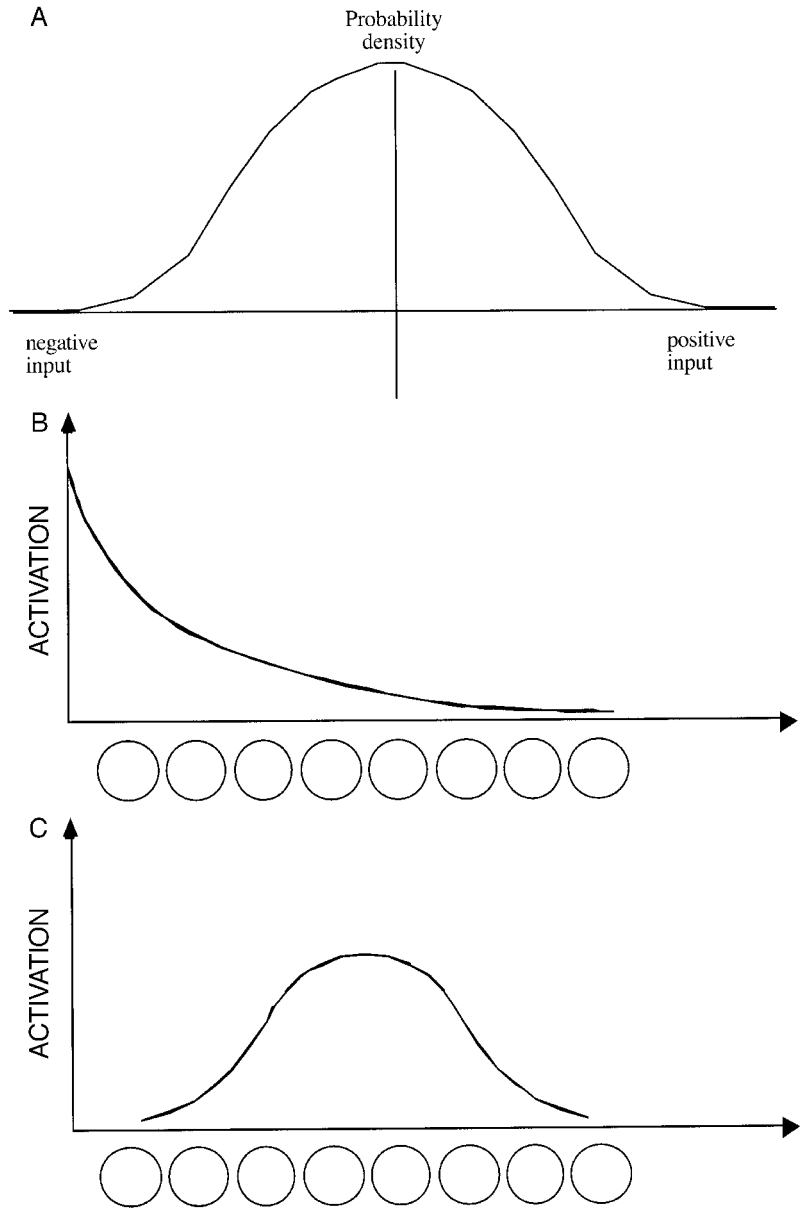

Figure 2. (A) An approximately Gaussian probability distribution for input to the units. (B) The distribution of activation across units that results from this input. (C) The desired distribution.

elements, new elements that were not present in the preceding stimuli (i, $i-1$, $i-2$, etc.). The choice of elements for deletion is subject to the constraint that they must be neither those in i that replaced some of the inputs in its preceding neighbor ($i-1$) nor inputs that replaced those of $i-1$, $i-2$, and so forth that were changed to give i. With these provisos, it will be possible to order the stimuli from one pole (extreme stimulus on the dimension) to the other; thus, they constitute a dimension.

As a concrete illustration of this proposal, consider a system with 100 potential perceptual elements, of which 20 are perceived and so provide input to our system at any one time. If we allow 40 representational units, each positively connected to half the inputs, the maximally active unit will be in receipt of approximately 15 inputs, or 5 more than would be expected by chance (which is $20 \times 0.5 = 10$). Now imagine that we change the set of active inputs 4 at a time. Changed inputs are not eligible for change, and we never reuse inputs that have been changed. For a unit with 15 active inputs, the probability of selecting an active input from its set of 20 is .75. Out of each four changes, we expect three to be active inputs,

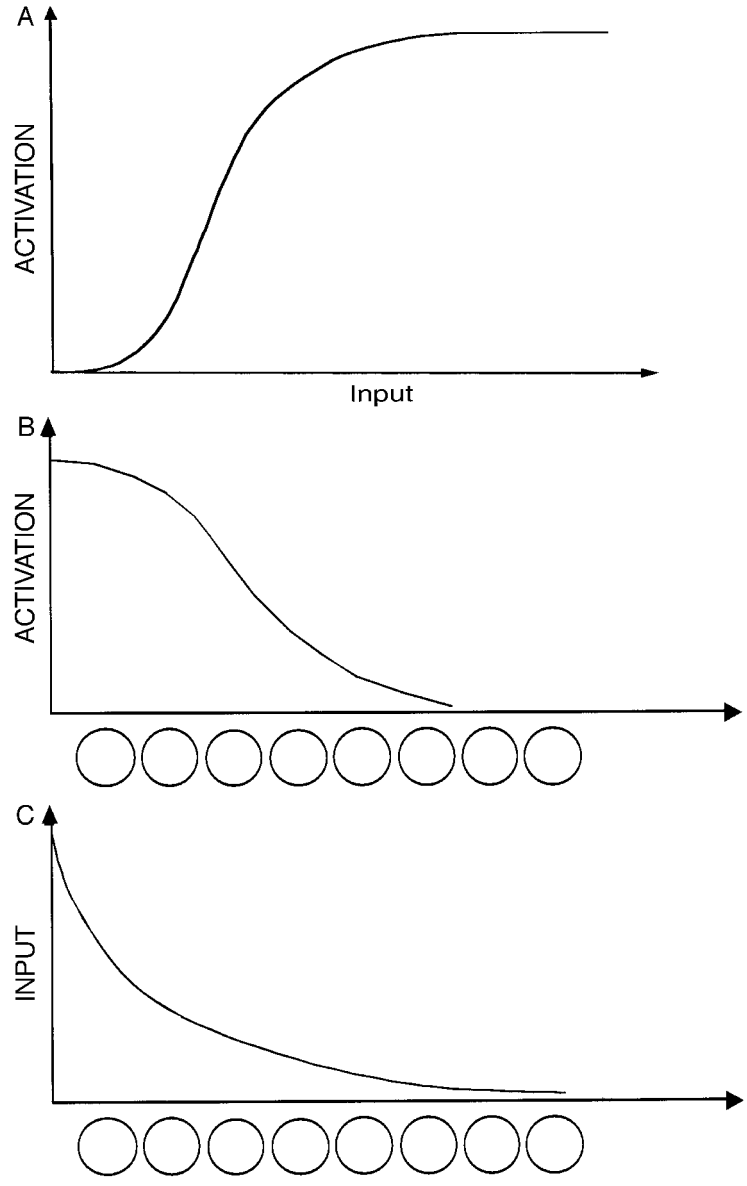

Figure 3. (A) The input–activation function for units in the system. (B) The activation distribution that results from combining the function shown in panel A with the input distribution shown in panel C.

and the new inputs will have, on average, 2 active inputs. Thus, there will be a net change of a reduction of 1 active input. Consequently, the more active units will, on average, decline in activity as we change the stimulus. At the same time, others will increase in activity, and the net result will be a smooth change from the initial stimulus to another via a series of intermediate steps—that is, a dimension.

This is far from being a complete account of our analysis. But it is sufficient for understanding how our model explains generalization and certain phenomena, such as the peak shift, that have been explained by appealing to interacting gradients of generalization. We shall defer to the section on discrimination learning our analysis of the representation of compound stimuli, but one final point about generalization needs to be addressed here. Given the assumption that there is considerable representational overlap for stimuli that have no special relation to one another, such as a tone and a light, why will there not be inappropriate generalization of responding from one to the other? The solution follows from our ability to specify what the expected overlap between two randomly chosen stimuli would be and then to specify a response function relating responding to asso-

ciative strength that gives little response for this fraction of λ. Thus, generalization may well occur but may not manifest in responding.

APPLICATION TO GENERALIZATION

Pavlov was the first to undertake systematic studies of stimulus generalization, reporting that "if a tone of 1000 Hz is established as a conditioned stimulus, many other tones spontaneously acquire similar properties, such properties diminishing proportionally to the intervals of these tones from the one of 1000 Hz" (Pavlov, 1927, p. 113). Similar gradients of generalization were reported for tactile stimuli as a function of the distance between the CS and the test stimulus. Pavlov interpreted such generalization as evidence for his concept of irradiation: Excitation of one particular cortical point was said to spread out, or irradiate, across the cortex, diminishing as it traveled further. It was left to Konorski (1948) to apply an elemental analysis. He argued that generalization was not a consequence of the irradiation of excitation from a single point of application, but of the distributed nature of the activation generated by the reinforced CS: "It must be assumed that the cortical centres of particular stimuli represent complex and widely dispersed formations, that they can partially overlap, and that this partial overlapping is the cause of generalization" (Konorski, 1948, p. 128). Subsequent physiological analysis has, here as elsewhere, strongly supported Konorski's analysis over Pavlov's (Thompson, 1965).

Hull (1943, p. 191) briefly toyed with the idea that stimulus generalization arose from the distributed nature of stimulus representations and the resulting overlap in the representations of similar stimuli. The idea was developed more formally in the statistical learning theory of Bush and Mosteller (1951) and stimulus sampling theory (Atkinson & Estes, 1963), and it was incorporated without much further comment into the Rescorla–Wagner model (Rescorla & Wagner, 1972) and into SOP (Wagner, 1981). Plausible and attractive as this analysis may be, however, one reason for Hull's caution was that he saw "no immediate prospect of securing a critical test of the hypothesis" (*ibid.*). How might one bring behavioral evidence to bear on the proposition that generalization of conditioned responding from one stimulus to another arises because the representations of the two stimuli are distributed across a number of "units" and, thus, on the supposition that conditioning of responding to one stimulus generalizes to the other by virtue of the responding that was directly conditioned to the elements or features they share in common?

One way to address this question is to study generalization between deliberately constructed stimulus compounds, which contain independently manipulable common elements or components. If two component stimuli, A and B, share relatively few elements in common, there will be little generalization from one to the other. But the addition of a specific common component, X, to form two compound stimuli, AX and BX, should result in substantial generalization from one compound to the other. Using elementary flavors (saline, sucrose, and lemon) as A, B, and X, Mackintosh, Kaye, and Bennett (1991) have confirmed these simple predictions. An aversion conditioned to A (saline) did not generalize to B (sucrose), but one conditioned to AX (saline–lemon) generalized strongly to BX (sucrose–lemon).

This generalization from AX to BX is dependent on the strength of conditioning to X. After the conditioning of an aversion to AX, a single extinction trial to X was sufficient to reduce the generalized aversion to BX, although a similar unreinforced trial to B had no effect on the aversion to BX (Bennett, Wills, Wells, & Mackintosh, 1994). Generalization from AX to BX can also be reduced by unreinforced exposure to X prior to training of AX, (Bennett et al., 1994; Mackintosh et al., 1991). Such preexposure causes latent inhibition to X and, when AX is poisoned, ensures that the aversion is conditioned preferentially to A, rather than to X. Finally, this explains why prior unreinforced exposure to BX itself is also sufficient to reduce generalization from AX to BX: Bennett et al. showed that the critical factor here is that such exposure results in latent inhibition of X (rather than of B).

In an ingenious set of experiments, Rescorla (1976) demonstrated some similar effects in the generalization of conditioning from one auditory stimulus to another. Following conditioning to a high-frequency (1800-Hz) tone, rats were tested for generalization to a low-frequency (350-Hz) tone. Animals that had received unreinforced preexposure to the low tone showed significantly less generalization than did animals that had received no preexposure. Rescorla argued that the two auditory stimuli should be conceptualized as AX (high) and BX (low), because they must be assumed to share elements in common, and that unreinforced preexposure to BX reduced generalization, because it caused latent inhibition of the common X elements.

In support of this analysis, in the next stage of the experiment animals received conditioning trials to BX and were finally tested with AX again. Animals that had initially received unreinforced preexposure to BX now showed *more* generalization from BX back to AX. Rescorla's analysis ran as follows. Latent inhibition of BX ensured that reinforcement of AX produced more conditioning to A than to X. Hence, there was little generalization to BX. But when BX was now reinforced, there was a substantial increase in the associative value of both B and X, and this new conditioning to X generalized back to AX. In the absence of this prior latent inhibition to BX in the control group, conditioning of AX conditioned X as well as A, and there was less room for reinforcement of BX to increase the value of X and so promote responding to AX. Rescorla's experiments provide convincing evidence that stimuli varying along a di-

mension should be conceptualized as sets of overlapping elements.

The Effects of Discrimination on Generalization: Peak Shift and Transfer Along a Continuum

Generalization between stimuli varying along a dimension such as wavelength or auditory frequency is, unsurprisingly, reduced by reinforcement of responding to one value on the dimension and nonreinforcement of responding to another value (e.g., Jenkins & Harrison, 1962). But such discrimination training has additional effects of considerable theoretical importance: the peak shift and transfer along a continuum, both of which have been explained by postulating interacting gradients of generalization of excitation around S+ and inhibition around S−.

The peak shift was first demonstrated by Hanson (1959). He trained one group of pigeons on a wavelength discrimination between an S+ of 550 nm and an S− of 560 nm. A control group was trained to respond only to the S+ wavelength. Both groups were then tested for generalization of responding to other wavelengths ranging from 460 to 620 nm. The discrimination group showed a steeper gradient of generalization than did the control, especially to longer wavelengths, at which they hardly responded at all; more important, however, although the peak of the control group's gradient was centered on their S+—that is, at 550 nm—that of the discrimination group was not; the latter responded significantly more to 540 and 530 nm than to 550 nm. The peak shift has been replicated in numerous other studies, in other animals and with other stimulus dimensions (see Honig & Urcuioli, 1981, and Rilling, 1977, for reviews). These results seem to confirm an analysis that had earlier been applied to the phenomenon of transposition by Spence (1937). Applied to the peak shift, Spence's analysis was as follows. Excitation conditioned to S+ generalizes to neighboring stimuli in the manner shown in Figure 4A, whereas inhibition conditioned to S− also generalizes to neighboring stimuli in a similar manner. The net excitatory value of any stimulus is given by subtracting inhibition from excitation, and it is clear that the net value of the stimuli labeled N+ and F+ in Figure 4A (near to and further from S+) will be greater than that of S+ itself. Note also, however, that as one progresses yet further from S+, the net value of the stimulus labeled FF+ is less than that of S+. This provides an accurate account of Hanson's data.

Transfer along a continuum was first demonstrated by Pavlov (1927), who found that if he wanted to train his dogs on a very difficult discrimination between a circle and a nearly circular ellipse, it was more efficient to train them, not on that discrimination from the outset, but ini-

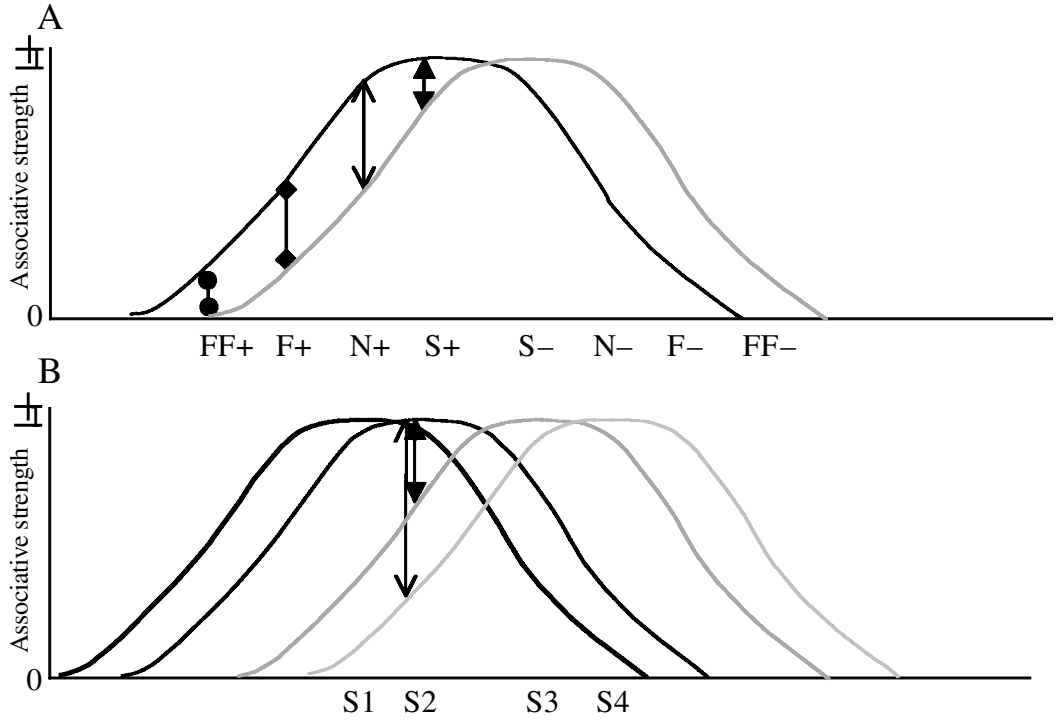

Figure 4. (A) An explanation of peak shift. The gray line shows the distribution of inhibitory associative strength; the black line shows the distribution for excitation. The solid double-headed arrow shows the difference between these distributions at S+, the open one shows the difference at N+, the diamond arrow shows the difference at F+, and the circle arrow shows that at FF+. (B) A similar analysis for transfer along a continuum. The differences shown are for the S+ on the hard discrimination. The open-headed arrow refers to the two gradients obtained with easy training, and the solid one refers to those derived from training on the hard problem.

tially on an easier version of the problem, between a circle and a long thin ellipse, and only gradually to transfer them to harder versions of the problem. Pavlov's results were confirmed by Lawrence (1952) in a study of brightness discrimination learning by rats and have since been replicated in several different species with a variety of different stimulus dimensions (see Mackintosh, 1974, for a review). Once again, as Lawrence (1955) and Logan (1966) have shown, the result can be explained by appeal to interacting gradients of generalization. The analysis makes use of the same gradients employed to explain peak shift. If, as is shown in Figure 4B, generalization gradients fall off more steeply as one moves further away from S+ and S−, differential reinforcement of two stimuli far apart on a dimension, S1 and S4, will result in a greater difference in the net excitatory value of two stimuli closer together on that dimension, S2 and S3, than will a comparable amount of differential reinforcement of S2 and S3 themselves.

It is not difficult to see that the shape of the gradients illustrated in Figure 4 is critical. Different functions will yield neither a peak shift nor transfer along a continuum. Linear functions predict equal net excitatory value for S+, N+, and F+. And exponential functions, falling off more slowly as one moves away from S+ and S−, predict that N+ and F+ will have *less* excitatory value than S+. Exactly the same predictions follow for transfer along a continuum. It might seem, then, that we have nicely converging, albeit indirect, evidence that generalization gradients must be of approximately the Gaussian shape shown in Figure 4. It is not, however, an easy matter to ascertain the "true" underlying shape of generalization gradients. Empirically obtained gradients of rate of responding to stimuli varying along a dimension will tell us what the true shape of the gradient is only if we know that there is an equal psychological distance between each stimulus along the dimension being assayed and have ascertained the relationship between our measure of responding and the underlying associative value of each stimulus. The careful analysis undertaken by Shepard (1987) concluded that the universal shape is exponential, and it is also worth noting that Pearce's (1987, 1994) configural theory assumes that generalization gradients are exponential. Our analysis implies that gradients are initially Gaussian but may become exponential after sufficient exposure to one or more stimuli on the dimension.

Regardless of this issue, however, there is a further reason to question whether the peak shift and transfer along a continuum can be explained solely by appeal to interacting gradients of the form shown in Figure 4. Empirical attempts to do so have been far from successful. The procedure employed in several studies of the peak shift has been as follows. An experimental group is trained on an intradimensional discrimination between S1+ and S2− (say, two wavelengths), and their postdiscrimination gradient is measured. Two further groups are trained on interdimensional discriminations, one between S1+ and S0− (where S0 is an achromatic stimulus) and the second between S0+ and S2−. The first of these two groups then provides an excitatory postdiscrimination gradient along the wavelength dimension centered on S1, whereas the second provides an inhibitory postdiscrimination gradient centered on S2. The latter group's inhibitory gradient can then be subtracted from the former group's excitatory gradient and the resultant can be compared with that obtained from the group trained on the intradimensional discrimination. Three such experiments have been reported (Hearst, 1968; Klein & Rilling, 1974; Marsh, 1972). Both Hearst and Rilling and Klein obtained reasonably close agreement between their observed and derived postdiscrimination gradients, but only after raw response rates had been transformed. Unfortunately, neither study observed any peak shift. Marsh did observe a peak shift in his experimental group, and the combination of the excitatory and inhibitory gradients generated by the two interdimensional groups did indeed predict a peak shift. But, as Marsh acknowledged, this predicted peak shift was notably smaller than that actually observed in the intradimensional group.

Mackintosh and Little (1970) performed an analogous experiment on transfer along a continuum. They first confirmed that transfer along a continuum would occur in pigeons trained on wavelength discriminations. One group of pigeons was trained on an easy wavelength discrimination between stimuli that we can label as S1+ and S4−; they then performed significantly more accurately when shifted to a harder wavelength problem (S2+ vs. S3−) than did birds trained on the hard problem from the outset. They then trained four further groups of birds on interdimensional discriminations between one of these four wavelengths and a plain white keylight (S0): One group was trained with S1+, one with S2+, one with S3−, and one with S4−. From Figure 4B, it is easy to see that the gradients of generalization of excitation and inhibition conditioned to S1+ and S4− should yield a greater difference in the net excitatory values of S2 and S3 than would direct reinforcement and nonreinforcement of S2 and S3 themselves. It follows, therefore, that the interdimensional groups trained with S1 and S4 should perform more accurately on the S2–S3 discrimination than those trained with S2 and S3. In fact, they performed rather *less* accurately.

Elemental Analysis of Peak Shift

Does this imply that peak shift and transfer along a continuum cannot be explained by interacting gradients of generalization? Not necessarily. The solution to the conundrum is to see that a truly elemental analysis of generalization and discrimination, when combined with an error-correcting learning rule, can yield quite different predictions from Spence's (1937) analysis.

Blough (1975) showed how an elementary analysis of generalization would predict the peak shift. A drastically

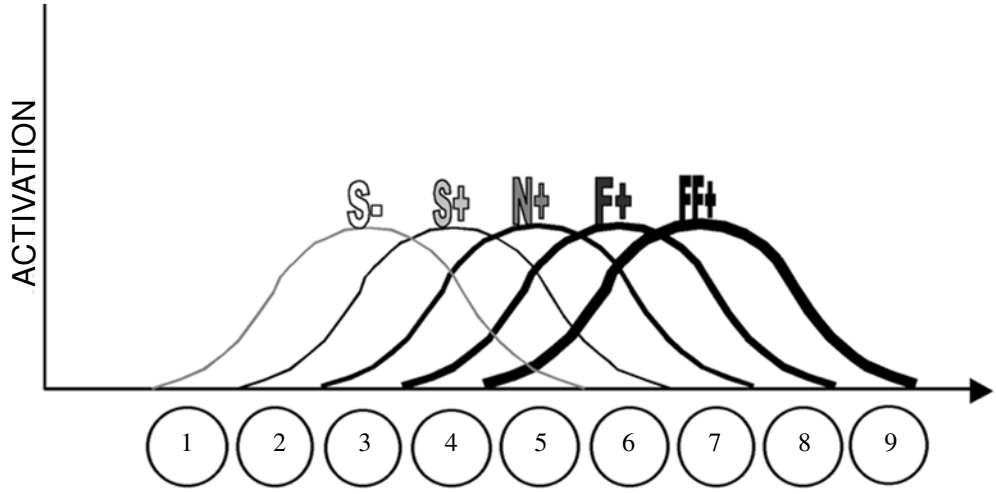

Figure 5. The dimensional analysis applicable to the peak shift experiments of Wills and Mackintosh (1998). See the text for details.

oversimplified version of this type of analysis is illustrated in Figure 5. A series of five stimuli are represented by activation of overlapping sets of units. The negative stimulus, S−, is represented primarily by activation of Unit 3, but also, to a lesser extent, by activation of Units 2 and 4 and, to an even lesser extent, by activation of Units 1 and 5. The neighboring stimulus, S+, primarily activates Unit 4, but also Units 2 to 6. And so on. If differential reinforcement is given for responding to S+, but not to S−, an error-correcting learning rule will ensure that excitatory conditioning accrues not to all units representing elements of S+, nor even necessarily to those most strongly activated by S+, but more selectively to those that differentiate S+ from S−. Units representing elements common to both will acquire relatively little excitatory value, unless they are much more strongly represented in S+ than in S−. The implication is that Units 5 and 6 will acquire at least as much excitatory value as Unit 4, which is activated fairly strongly by both S+ and S−. If, however, animals are trained on an interdimensional discrimination between this S+ and an S0 that activated none of the units shown schematically in Figure 5, the pattern of weight changes to the units activated by S+ would simply be proportional to their level of activation, with Unit 4 changing most, followed by Units 3 and 5, and so on.

The critical point is that intradimensional discrimination training between S+ and S− will yield a pattern of weight changes to the units representing S+ quite different from that produced by interdimensional discrimination training between S+ and S0. There is, thus, no reason to expect that the peak shift, or transfer along a continuum, will be well modeled by the gradients of generalization resulting from such interdimensional training.

The remaining three stimuli illustrated in Figure 5 are, of course, test stimuli near to (N+), far from (F+), and yet further from (FF+) S+. Since stimulus N+ activates Units 5 and 6 more strongly than S+ (and Units 2 and 3 less strongly), it is easy to see that it should command a higher rate of responding than S+. Whether F+ would have a higher net excitatory value than S+ is debatable, but it seems clear that FF+ must have less. Hence, the analysis predicts that the peak of responding will be to N+ and that responding will decline as one moves to F+ and FF+.

Figure 5, in fact, provides a schematic representation of the stimuli used by Wills and Mackintosh (1998) to test this elementary analysis. They constructed a series of stimuli varying along an artificial dimension. Each stimulus consisted of a 3 × 4 array of arbitrary icons—for example (see Figure 6), four instances of a "central" icon, three instances of two others, and one instance of each of two others (in fact, the precise construction of the stimuli varied from one experiment to another without materially affecting the outcome). The most important difference between Figure 5 and the stimuli used by Wills and Mackintosh was that their icons were wholly arbitrary: The icons did not themselves form a series, as the numbered units in Figure 5 do. Nevertheless, the stimuli constructed from these arbitrary icons did form a series, as the results of their experiments showed. Pigeons and people learned to discriminate between neighboring stimuli (S+ and S− in Figure 5) and, when tested with near and far stimuli, showed both positive and negative peak shifts, responding more to N+ than to either S+ or FF+ and less to N− than to S− or FF−. The results for the pigeons are shown in the top panel of Figure 7.

The reason for this peak shift was that, as an error-correcting rule would predict, responding was primarily controlled by those elements or icons that served to discriminate between S+ and S−. Following the peak shift test, the pigeons were given a further set of test trials to

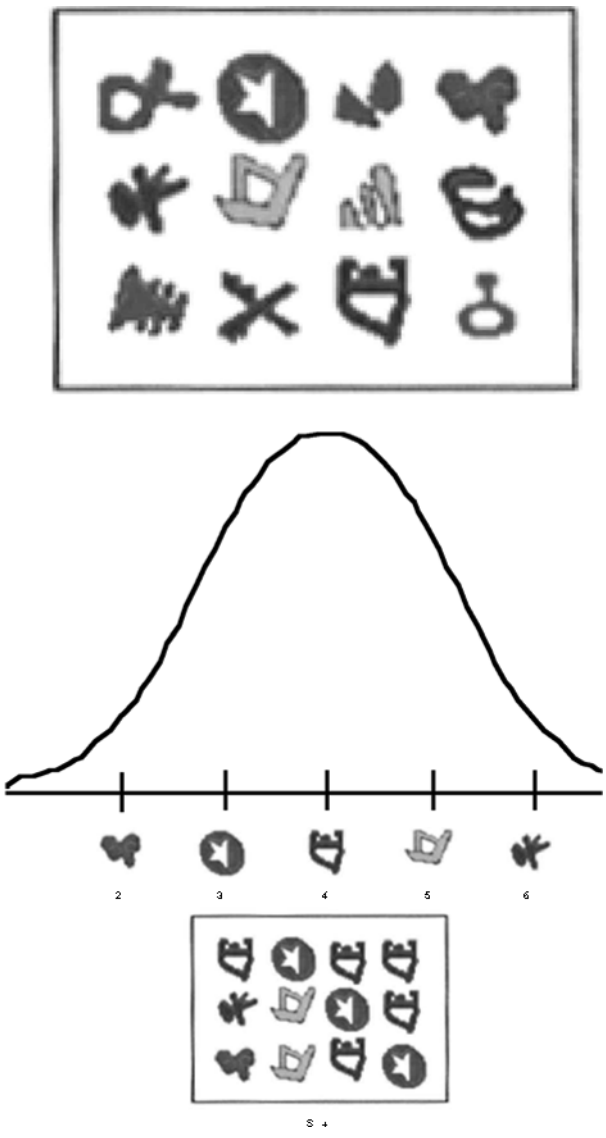

Figure 6. The artificial icon-based stimuli and the algorithm used for generating them. Icons are chosen from the set illustrated at the top of the figure according to the dimensional value being represented. The type of distribution used for a given dimensional value is illustrated immediately beneath the icon set, and the resultant stimulus is shown at the bottom of the figure.

stimuli that contained several copies of one of the icons present in either S+ or S−, together with some filler icons that they had not seen before. Recall from Figures 5 and 6 that S+ consisted of Icons 2, 3, 4, 5, and 6, with more of Icon 4 than of 3 or 5 and more of 3 and 5 than of 2 or 6. Although Icon 4 was thus the most frequently occurring element of S+, it also occurred quite frequently in S−, and as can be seen in the bottom panel of Figure 7, the results of this test showed that it was Icon 5, not 4, that commanded the highest rate of responding, whereas the comparable icon of S−, number 2, commanded the lowest rate.

Figure 7 also shows the results predicted both for the peak shift test and for the individual icon test by a simple error-correcting model. The fit between model and data was good. These results have been confirmed in humans trained with Gaussian distributions (Jones & McLaren, 1999), whereas Wills and Mackintosh (1998) also found a peak shift when pigeons were trained with triangular distributions (3, 6, 3). This last result has been extended by Oakeshott and Mackintosh (2002). They found that pigeons trained to discriminate between stimuli consisting of either triangular or quasi-Gaussian distributions of icons produced a peak shift on test but, importantly, two other distributions did not yield a peak shift. One was a sharply exponential distribution (1–3–7–3–1); the other was a rectangular distribution. In the former case, the discrimination between S+ and S− would be based largely on the central icon, the seven copies of which made up nearly half of the 15 icons present in each stimulus. Since N+ had only three copies of S+'s central icon, it is not surprising that it failed to command a higher rate of responding than did S+. And in the case of a rectangular distribution, the elemental analysis has no basis for predicting a peak shift. If each of 6 icons appears twice in a 12-icon stimulus, the discrimination between S+ and S− is based on the two instances of the icon unique to S+ and the two instances of the icon unique to S−, with the other icons present being common to both stimuli. The two instances of the unique S+ icon will indeed be represented in N+, but the two instances of the new icon in N+, being wholly novel, can have no associative value, and there is therefore no basis for expecting more responding to N+ than to S+. Single icon testing confirmed that the 5 icons common to S+ and S− commanded a low rate of responding; the icon unique to S− commanded an even lower rate, whereas the icon unique to S+ commanded a high rate.

This analysis makes it clear that something like the distributed patterns of activation over representational units that we are able to generate for stimuli in our model will be necessary to generate appropriate peak shift effects. The reason that icon-generated stimuli result in this type of representation under the conditions in force in these experiments is that the random rearrangements of the constituent icons that take place from trial to trial ensure that the *core* units (defined later on p. 192) representing each icon are the only ones relevant for learning, and so the predictions of *classical* elemental theory, as outlined above, follow.

We do not believe that a configural theory can easily account for these data, even if it incorporates a generalization rule different from that adopted by Pearce (1987, 1994). Indeed, when a configural solution to these multiple-icon discriminations is encouraged, no peak shift is observed. In all of the above experiments, the position of individual icons in each stimulus varied randomly from trial to trial. Oakeshott and Mackintosh (2002) compared this condition with one in which the position of each icon in a 12-icon stimulus remained the same on all trials. This latter condition, unlike the former, yielded no peak shift.

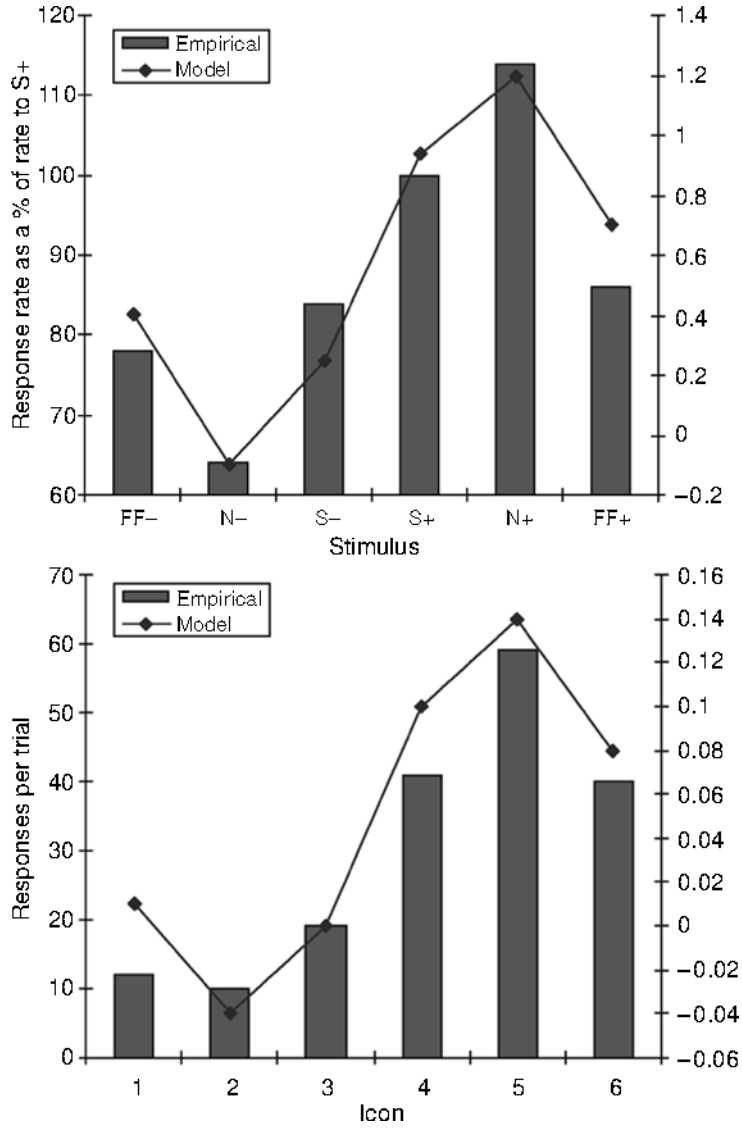

Figure 7. Peak shift results from an experiment by Wills and Mackintosh (1998) with the icon-based stimuli and results from a model implemented along lines set out by Blough (1975). The interval between training near and far stimuli was one unit on the dimension shown in Figure 6. The top panel shows the obtained and predicted results for the peak shift test; the bottom panel shows the obtained and predicted results for the tests with individual icons. The right *y*-axis in each graph gives the scale for the associative strengths generated by the model.

The simplest interpretation of this finding is surely that the fixed configuration of icons encouraged a configural solution that, as Pearce's theory predicts, did not yield a peak shift. Our model brings two factors to bear that support this analysis. First, the stable stimulus becomes rapidly familiar, which will tend to lead to a more exponential pattern of activation representing the stimulus; second, the emphasis on icon configurations means that the same changes in the icons used in earlier experiments now denote much greater changes in the stimulus representations. Consider as an illustration of this latter point, the case in which 3 out of a total of 9 icons are changed. On an icon by icon basis, this would be a change of 33%, leaving 67% of the original stimulus. If, instead, we consider configurations, taking pairs of icons as an example, the change is almost twice as much, leaving only about 42% of the original stimulus configuration. Thus, a change that might have revealed a peak shift when individual icons governed responding might now just show a drop in responding, because of the larger deviation from the original stimulus.

Pearce's (1987) configural theory took no account of the spatial arrangement of elements in a stimulus array; thus, random variation of the position of individual icons

is ignored. That this cannot be correct is suggested by the above results: There *is* a difference between fixed and varying arrays of icons. And George, Ward-Robinson, and Pearce (2001) have now shown, in an elegant set of experiments, that pigeons do indeed discriminate between different spatial arrays of the same set of elements. But can they really be supposed to have learned our standard multiple-icon discriminations by associating each of the several hundred different spatial configurations of S+ icons with reinforcement and the equal number of S− spatial configurations with its absence? And if they had, how would a peak shift have occurred? Application of Pearce's theory to the many configurations would still generate the same generalization gradients as those for the case of fixed spatial arrays when icons were changed and, so, would lead to the same predictions with regard to peak shift. This would be hard to reconcile with the quite different results obtained with the two types of array. It seems more plausible to argue that the random spatial arrays forced an elemental solution to the original discriminations and then to accept the elemental analysis of the peak shift.

Elemental Analysis of Transfer Along a Continuum

Let us now return to Mackintosh and Little's (1970) demonstration of transfer along a continuum and their failure to confirm the predictions of a gradient interaction account of the effect following interdimensional discrimination training (S+ vs. S0− and S− vs. S0+ separately). Our preceding analysis of the peak shift explains why interdimensional training may have a different impact on generalization than does intradimensional training, and this is equally critical for the analysis of transfer along a continuum. But the reason that interdimensional training between S1 (or S4) and S0 does not produce better discrimination between S2 and S3 than does comparable training between S2 (or S3) and S0 is that, by the end of this training, the gradients of generalization will no longer be Gaussian. Recall from our earlier analysis that the representation of any stimulus starts out as a Gaussian distribution of activation across a set of units but that, as the stimulus becomes more familiar, salience modulation tends to produce a more exponential distribution.

The position will be different following intradimensional training between S1 and S4 or between S2 and S3, because now it will be the difference between the two patterns of activation (for S+ and S−) that governs learning. As was noted above, this means that it is not necessarily the units maximally activated by S+ that will acquire the strongest associative strength but, rather, those that are more distant from those activated by S−. Even if the pattern of activation across units has become more exponential by the time of testing, it is the pattern of associative strengths set up by the initially Gaussian activation distributions that will determine responding and produce the transfer along a continuum effect.

An Empirical Example

To conclude this section, we will describe the results of a series of experiments in which both peak shift and transfer along a continuum were examined simultaneously in human subjects. An example of the stimulus di-

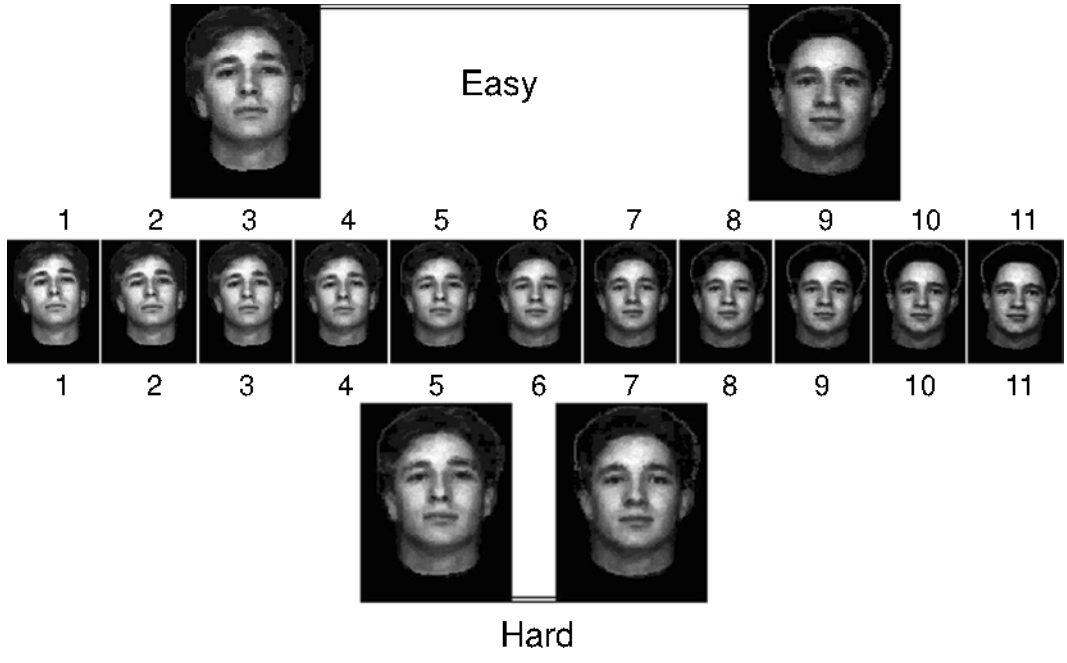

Figure 8. The morphed face dimension used in the experiments reported in this paper.

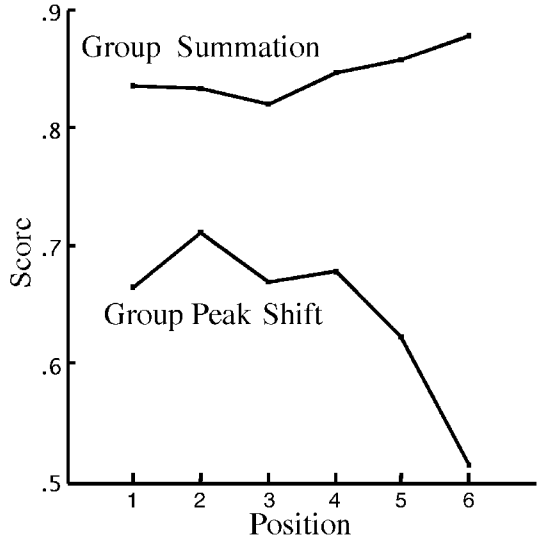

Figure 9. Peak shift and summed generalization on the morphed face dimension. The results shown here have the scores for Stimuli 1 and 11, 2 and 10, 3 and 9, and so on, averaged for those stimuli, as well as across subjects. This eliminates any key bias and allows a direct comparison of the two groups. The scores reflect the proportion of correct responses to the stimuli (chance = .5).

mension used is shown in Figure 8. It is constructed by taking pairs of faces and morphing from one to the other in 10 equal steps, giving a continuum with 11 points on it. The faces in a given pair were chosen to be similar (which aids the morphing process in keeping the transitions smooth) so that neighboring stimuli on the dimension are very similar indeed. Figure 8 illustrates the morphed face dimension for one pair of faces; the faces at 3 and 9 on the dimension always constituted the easy problem, and those at 5 and 7 the hard version. Four dimensions were used concurrently for every subject, with the assignment of the face dimensions to the conditions of the experiment counterbalanced appropriately. Pilot testing revealed that the discriminations used were difficult (even for the 3 vs. 9 case) but possible under the conditions of these experiments, and the subjects reported that their performance was hard to characterize in terms of rules based on features (desirable if performance is to be associatively driven).

Clearly, the algorithm for generation of the morphed face dimension fits well with the prescription for a dimension that we gave earlier. Starting at one pole (face) of the dimension, the algorithm proceeds by successively changing pixels of the image so as to arrive at the other pole (face). Once pixels are changed, they are not changed again. Given this, we expect the representation of this dimension to be as discussed on pages 179–180.

The demonstration of a peak shift involved training subjects to discriminate between Stimuli 3 (S+) and 9 (S−) and testing responding to stimuli across the dimension without giving the feedback used in training. The results, shown in Figure 9, revealed significantly more responding to Stimulus 2 than to 3 or 1, a classic peak shift.

A summed generalization effect was demonstrated by training subjects with both Stimuli 3 and 9 reinforced and testing responding to the intermediate stimuli. As can be seen in Figure 9, the subjects responded significantly more to Stimulus 6 than to the average of 3 and 9. Finally, we measured transfer along a continuum by taking two groups, one pretrained on the relatively easy discrimination between 3 and 9 and the other on the harder discrimination between 5 and 7. When both groups were then trained and tested on the 5 versus 7 discrimination, the *hard* group performed less well than the *easy* group, as can be seen in Figure 10.

These results are important both in their own right and in combination with one another. Taking the summed generalization result first, on its own it suggests good generalization to Stimulus 6 as a result of the training at Positions 3 and 9. Taken together with the peak shift result, these two effects strongly suggest a Gaussian pattern of activation as the representation for stimuli on this dimension. This would allow explanation of the peak shift *and* produce sufficient generalization to allow responding at Position 6 to exceed that at Positions 3 and

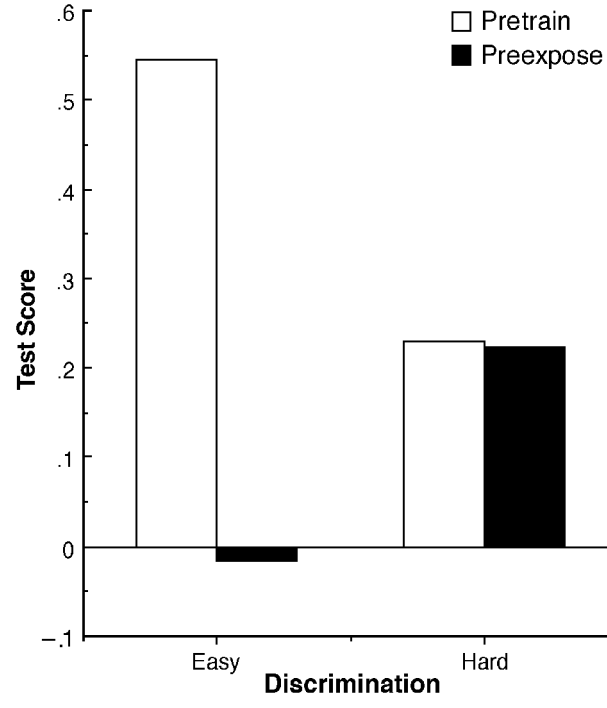

Figure 10. Transfer along a continuum: the effects of pretraining and preexposure. In this experiment, one key (e.g., the left key) was designated the negative category (a press scores −.5 for that stimulus), and the other key was designated the positive category (scores +.5) during test. The test score shown indicates the mean difference of the keypresses across subjects, would be zero if the subjects were indifferent to which stimulus went with a given key, and ranges from +.5 to −.5.

9 after training at 3 and 9. Exponential or power law gradients would not deliver sufficient generalization to enable this. As an illustration of this, consider how Pearce's configural theory would apply to these stimuli. If we make the simplifying assumption that we have stimuli only at the training and testing positions on our dimension and that these are represented as AXY at one training position and as BXZ at the other, with XYZ in the middle, and then apply Pearce's rules for generalization, we obtain the following: The representations of AXY and BXZ will both have associative strength of 0.9 lambda in their own right, so that the total generalized associative strength for each of the stimuli AXY and BXZ is lambda, and this implies that stimulus XYZ will have generalized associative strength of 0.8 lambda, which is not the result obtained.

It is difficult to see how this analysis could alter its prediction that responding to the middle stimulus will never be greater than responding to the trained stimuli. If we make the trained stimuli less similar—that is, AX and BY with the middle stimulus XY—then generalization drops to 0.5 lambda. If we increase the similarity of the two trained stimuli, then in the limit, generalization to the middle stimulus will be lambda, the same as that for the trained stimuli, but not higher.

Our own analysis of the pattern of effects obtained with transfer along a continuum is also quite straightforward. Transfer occurs because of the generalization already postulated to explain the summed generalization reported earlier. Generalization from 3 to 5 and from 9 to 7 will be good, but because of the Gaussian gradients involved, generalization from 3 to 7 and from 9 to 5 will be much poorer. This means that training on the easy discrimination will enable subjects transferred to the harder one to discriminate on the basis of generalized associative strengths straight away, and this will also facilitate further acquisition of the discrimination. Suret and McLaren (2002) have demonstrated that this is predicted by our model and that training on the hard discrimination from the outset is not as effective. Their model implements the theory given here and in McLaren and Mackintosh (2000), and the simulations accurately reproduce the pattern of effects reported in this section.

A further set of subjects in these experiments were split into two groups and simply exposed to the easy or hard stimuli, rather than being pretrained on them. The rationale for this was that there is another analysis of transfer along a continuum that does not require an elemental analysis and, instead, is more in the configural tradition. This approach states that the reason for good transfer is that training on an easy discrimination enables good configural representations of the S+ and S− stimuli to be developed and that these representations can then be used to solve the harder discrimination on that dimension. This account is not based on Pearce (1987; indeed, it is hard to derive on that basis) but is one offered by Saksida (1999). One feature of this account is that it predicts that mere exposure to the stimuli that constitute the easy discrimination should prove beneficial in the same way that explicit training on the easy discrimination helps when acquiring the more difficult one. McLaren and Suret (2000) were able to show that preexposure to the easy stimuli was less beneficial than preexposure to the hard stimuli, and that is also the case with the data reported here, casting further doubt on this type of configural analysis of these issues.

We now move our focus from generalization to a more thorough analysis of discrimination. Some notion of discrimination has, of course, been implicit in our discussion so far, but now we further develop our model to show explicitly how an elemental model of associative learning can account for discrimination phenomena.

THEORY OF DISCRIMINATION LEARNING

Elemental and Configural Theories of Discrimination

The most common approach to the analysis of discrimination learning, exemplified by the *conditioning–extinction* theories of Pavlov (1927), Spence (1936), and Hull (1952), has been to postulate changes in the associative strength (or related construct) of particular component stimuli—a red light, a circle, a white door, or the left arm of a maze. In differential conditioning or successive discrimination training, excitatory conditioning accrues to one component stimulus (S+; say, a red light) associated with reinforcement, while inhibitory conditioning conditioning accrues to another (S−; say, a green light) associated with its absence. Spence (1936) applied a similar analysis to the case of simultaneous discriminations, in which both S+ and S− appeared on the same trial, typically side by side, with each stimulus appearing equally often on the left and on the right. Spence showed how changes in the associative values of the four component stimuli (red, green, left, and right) could account for a variety of features of the learning of such discriminations. The attentional theory of Sutherland and Mackintosh (1971) applied a similar analysis.

According to another class of theory (e.g., Gulliksen & Wolfle, 1938), however, the two kinds of trials of a simultaneous discrimination involve the presentation of two distinct stimulus configurations, red to the left of green on some trials and green to the left of red on others, and the problem is solved by learning to make one response (left) to one configuration and another (right) to the other. Yet a third class of theory is possible (e.g., Medin, 1975): Animals might learn to respond to particular color–position compounds—for example, to red–left and red–right in preference to green–right and green–left.

Both Gulliksen and Wolfle's (1938) and Medin's (1975) theories can be regarded as configural theories. They differ, of course, in whether they see all four component stimuli (red, green, left, and right) as forming a configuration, or only two at a time (red and left or green

and right, etc.). But they can still be contrasted with the elemental or component approach of Spence (1936) and Sutherland and Mackintosh (1971). A review of a large earlier body of evidence (Mackintosh, 1974) suggests that all three approaches are needed to explain the results of all experiments. For example, Spence's original theory is unable to explain the solution of so-called *successive-conditional* discriminations, in which two red stimuli are presented on some trials and choice of the left one is reinforced, and two greens on other trials with choice of right reinforced. Each component (red, green, left, and right) is equally often reinforced and not reinforced, and the combination of any two should have the same total value as the combination of any other two components.

More recent studies have usually employed successive discrimination training, with S+ and S− presented on separate trials, in which the traditional component approach may seem more natural and simpler; but just like Spence's (1936) theory, a component account is unable, unaided, to explain all instances of successful discrimination. The two best-known failures are the negative patterning, or XOR (exclusive or), problem and (again) conditional discriminations. In a negative patterning discrimination (Woodbury, 1943), two stimuli, A and B, are reinforced when presented alone, but the AB compound is not; if the associative value of AB is simply the sum of the values of A and B, the compound will command a higher rate of responding than do its components. In a conditional discrimination (Saavedra, 1975), there are trials involving four compounds (AX, BX, AY, and BY) with AX and BY reinforced and BX and AY not reinforced. Just as in the red–left, green–right discrimination, each component stimulus is equally often reinforced and not reinforced, and a simple componential analysis predicts that the associative value of all four compounds will be the same.

Pearce (1987, 1994) has proposed a configural theory of conditioning and discrimination learning that, in a manner analogous to that of Medin's (1975) *compounding* solution of successive-conditional discriminations, explains the solution of such problems by proposing that excitatory conditioning and inhibitory conditioning accrue, not to the elements of a single stimulus or to the components of a compound stimulus, but to the entire configuration present on any trial. The negative patterning problem is solved because, although excitatory conditioning will accrue to both A+ and B+, the AB compound is not just the sum of A and B; it is a distinct stimulus in its own right, to which inhibitory conditioning can accrue. Similarly, Saavedra's (1975) conditional discrimination is solved by conditioning excitation to AX+ and BY+ and inhibition to BX− and AY−. Of course, Pearce acknowledges that AX and AY are similar to one another, as are BY and BX, and that AB is similar to A and B. There will, therefore, be generalization between them, making the discriminations harder than a simple discrimination between A and B. But the important point is that, in the negative patterning discrimination, in which both A and B are reinforced, any responding to the AB compound is simply a consequence of generalization from A and B, and AB itself is available to be associated with the absence of reinforcement. Pearce distinguishes between the excitation or inhibition that is directly conditioned to any stimulus as a result of its pairing with the presence or absence of reinforcement and the generalized excitation or inhibition that comes to it from other, similar stimuli. The total associative value of any stimulus is the sum of these two.

So far, so good. What might seem somewhat more surprising is that Pearce appeals to common elements to explain generalization. In the negative patterning problem, there is generalization from A and B to AB and from AB back to A and B, because AB shares elements in common with A and B. Pearce's rule for generalization is, indeed, quite straightforward. Assuming that A and B are equally salient, A and B each comprise half the elements of AB, so the generalization between them is .5. In the conditional discrimination, again assuming equal salience of the four component stimuli, the generalization between those compounds that share a common component is .25: the generalization between AX and BX is dependent on the X component they share in common, which comprises half the elements of each. Formally, Pearce's rule is

$$ASA' = \frac{P_{com}}{PA} \cdot \frac{P_{com}}{PA'}, \quad (1)$$

where ASA' is the similarity between, and therefore the generalization between, Stimuli A and A′, P_{com} is the number (strictly total salience) of the elements they share in common, and PA and PA' are the total saliences of all the elements in A and A′ respectively.

Our first point with regard to this approach to computing similarity is that it would be fairer to characterize it as a configural model built onto an elemental one, given the vital role that an elemental level of representation plays in determining generalization coefficients. Our second point is that it is quite possible for an elemental model of the type that we envisage to accommodate the phenomena often taken as evidence for configural representation. As Wagner and Brandon (2001) have shown and as we argue below, there are two paths that an elemental account can follow in order to explain negative patterning and conditional discrimination learning. The first assumes that when two component stimuli, A and B, are presented together, there are new elements generated by the AB compound that were not present in either A or B alone. This is the *unique cue* solution to negative patterning proposed by Wagner and Rescorla (1972) and Rescorla (1972). Wagner and Brandon refer to this as an *added element* approach. The second assumes that when A and B are presented together, some of their elements are *not* represented in the AB compound. This possibility was noted by Rescorla and Wagner in their original paper (1972, p. 86), and Wagner and Brandon (2001) have shown that this model (their inhibited elements

model) is, to a first approximation, equivalent to Pearce's configural model. The first assumption can be implemented by saying that the AB compound should be conceived of as ABX (X being the unique cue) and, in Rescorla–Wagner terminology, the negative patterning discrimination is solved when A = B = λ and X = -2λ. The second can be implemented by saying that A and B should be conceived of as AX and BX, but the compound as ABX, rather than ABXX. Then, the problem is solved when X = 2λ and A = B = $-\lambda$.

Wagner and Brandon (2001) incorporated both of these assumptions into their replaced elements model. As we shall presently argue, our own elemental analysis of stimulus representation leads to an account of discrimination learning that, in some circumstances, is very similar to Wagner and Brandon's replaced elements model. But we begin by noting an aspect of our model that we believe may have no ready equivalent in theirs. We are still assuming, of course, that stimuli are represented as sets of elements and that these elements generate patterns of activation distributed over a set of units. These units will vary in the level of activity in response to a given stimulus, but they also vary in the extent to which their activation depends on other stimuli present at the time. We define *core* units as ones that will be activated mainly by inputs corresponding to elements of the stimulus itself and whose activation will be relatively independent of other stimuli present at the time. But other, *peripheral* units will be activated by a combination of elements across different stimuli, and their level of activation will be partly, and in some cases largely, determined by whatever other stimuli happen to be present at the same time. It is these latter units that will be sensitive to changes in context or accompanying stimuli; they bear some similarity, of course, to those employed in some configural theories of stimulus representation.

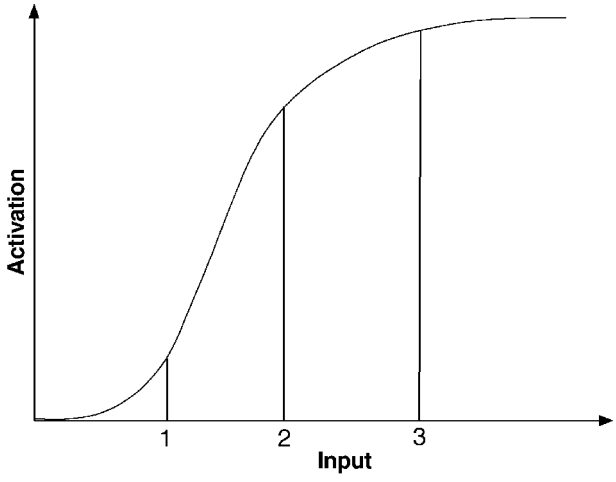

Figure 11. The function relating input to activation for a unit. Thus, if a unit receives the input from two stimuli (separately), shown at 1, then in compound, these stimuli will activate the unit far more than any linear sum of their individual effects would imply (shown at 2), resulting in a unit that can function as a unique cue for the compound. If both stimuli activate the unit as shown at 2, the compound will activate the unit little more than either of the individual stimuli (shown at 3), and the unit will act as an instance of a nonadditive common cue.

Implications for the Representation of Compound Stimuli

The analysis outlined above also allows us to offer a more principled discussion of the way in which compound stimuli are dealt with in our theory. When two component stimuli, A and B, are presented in a compound, how will that AB compound be represented? A very simple-minded elemental theory might assume that we just add together the input from the elements of A with the input from the elements of B. But our account implies two qualifications to this simple story. Following Konorski (1948), we have already assumed that our representational units have input–activation functions as shown in Figure 11 and that representations are distributed so that any given unit takes part in the representation of many different stimuli. If A and B each individually activates a unit as shown at Point 1 on the input axis, the compound of the two stimuli will result in something like the response at Point 2. But if each of A and B produces an input to the unit corresponding to Point 2, the AB compound will produce a response as at Point 3—that is, only a small additional impact on the activity of this unit, relative to the activity produced by its components. The former case, in which the AB compound results in a level of activation substantially greater than the sum of those from A and B separately, corresponds to the idea that compounds generate *unique cues* (Rescorla, 1972, 1973; Wagner & Rescorla, 1972). The latter case, in which the compound results in a level of activation substantially less than the sum of those from A and B separately, corresponds to the idea that the elements common to two component stimuli are not added together when the two stimuli are compounded (Rescorla, 1972; Rescorla & Wagner, 1972).

One consequence of the distributed representational scheme that we have adopted is that stimuli from different modalities, which we might think of as having few, if any, features in common (e.g., a tone and a light), might well have considerable overlap in representational terms. Thus, this analysis will apply not only to the case in which A and B are similar stimuli from the same modality. It may also hold (although, no doubt, to a lesser extent) when we are dealing with quite different stimuli from different modalities. We are not, of course, saying that a tone and a light are similar but only that there will be a certain baseline level of overlap even for stimuli that are quite dissimilar. Obviously similar stimuli will have still greater degrees of representation overlap.

This account is obviously very similar to Wagner and Brandon's (2001) replaced elements model. As we have noted in Figure 11, the level of activation of units responding to A and B alone may be more than twice as

great when the AB compound is presented. This provides the basis for postulating added elements or unique cues. We also allow for inhibited elements, since a unit activated by A alone may become inactive when B is added to A, because the change in the pattern of inputs may be such as to cause the net input to this unit to drop to zero. Given these similarities, our account will often yield the same predictions as Wagner and Brandon's. We would argue, however, that our account provides a more fundamental rationale for their assumptions than they do, and there are features of our account that find no immediate parallel in theirs. As we have noted, another reason that the elements present in component stimuli are underrepresented in the stimulus compound, in our account, is that the elements common to two component stimuli will not be represented twice over in the compound. This mechanism is distinct from the inhibited element mechanism noted above. The key here is that whereas in Wagner and Brandon's model, common elements may well behave in a nonadditive fashion, so will other elements. The model makes no distinction between common and unique elements insofar as their replacement when the stimuli are compounded is concerned. Moreover, the representation of stimuli and stimulus dimensions as distributed patterns of activation across sets of units is a feature of our model that has no ready equivalent in theirs. This follows partly because Wagner and Brandon did not seek to give a treatment of dimensions as such and partly because our distinction between *core* and *peripheral* units (see p. 192) is one that has no ready analogue in their model.

APPLICATION TO DISCRIMINATION

Common Elements and Learning per Trial

From the analysis just given, it seems intuitively obvious that an elemental account of this nature will predict that speed of discrimination learning between two stimuli will be inversely related to the proportion of elements they share in common, since the presence of such elements cannot contribute to successful discrimination. What seems intuitively obvious, however, may not always be true. Pearce (1987) has shown, by simulation, that one simple elemental model (Rescorla & Wagner, 1972) often predicts that an increase in the number of common elements shared by two stimuli, so far from decreasing, actually increases speed of discrimination learning. For example, the model predicts that the discrimination between AX+ and AXY− will actually be learned *more* rapidly than that between A+ and AY−. Since the opposite outcome is observed, Pearce has argued that there is a fundamental flaw in elemental theories of discrimination.

Why does the Rescorla–Wagner (1972) model make this counterintuitive, and counterfactual, prediction for an AX+, AXY− discrimination? In effect, the problem arises from the model's strong assumption that the associative strengths of two components of a compound stimulus are simply added together to yield the associative strength of the compound. It follows that the positive associative strength of the compound AX will increase faster than that of the single stimulus, A. Recall that on the first conditioning trial, according to the Rescorla–Wagner model, there will be no overshadowing between A and X, so that, assuming they are of equal salience, the associative value of AX will be twice that of A after the first reinforced trial. It is only at asymptote that $V_{AX} = V_A$. Discriminative performance will, at any stage, be a function of the difference in associative value of reinforced and nonreinforced stimuli—that is, $V_{AX} - V_{AXY}$ in one case, and $V_A - V_{AY}$ in the other. It is clear that these differences will be a simple consequence of the magnitude of inhibitory conditioning to Y. This, of course, depends on the error term, $\lambda - V$, where $\lambda = 0$ (for nonreinforced trials), and V = the sum of the positive associative strengths of A and X, or of A alone. Since $V_{AX} > V_A$, it follows that inhibitory conditioning to Y will proceed more rapidly in animals trained on the AX+ versus AXY− discrimination than in those trained on the A+ versus AY− discrimination. Pearce (1987) has shown that adding unique, configural cues (cf. Rescorla, 1972) will not alter this outcome.

One solution to this problem for a model such as Rescorla and Wagner's (1972), then, is to relax its strong summation rule and suggest that the associative value of a compound is *not* the arithmetic sum of the associative values of its components. We return to this possibility below. First, however, we shall argue that a real-time elemental model, such as ours, is not constrained to predict the same outcome as Rescorla and Wagner. Even if we go along with the assumption of strict summation of associative strengths, the outcome predicted by our model for such experiments depends on the amount of learning per trial: When the learning that occurs on each trial is low, we predict the same outcome as Rescorla and Wagner, but when it is high, we predict that the addition of common elements will indeed slow down discrimination learning. Critically, this latter prediction depends on the real-time nature of our model.

In order to see how the model makes this prediction, it will help to work through an example. Let us assume that learning proceeds to asymptote on each trial (this is by no means necessary but makes the exposition a great deal clearer). Consider two trials of discrimination training, either AX+ followed by AXY− or A+ followed by AY−.

1. On the AX+ trial, both V_A and V_X are incremented by .5 (this because of the real-time feature of the model given in McLaren & Mackintosh, 2000, which predicts one-trial overshadowing).

2. On the AXY− trial, V_A, V_X, and V_Y are all decremented by .33. Thus, at the end of Trial 2, $V_A = V_X = .17$ and $V_Y = -.33$, so that $V_A + V_X = .33$ and $V_A + V_X + V_Y = 0$.

Now consider the case in which a trial to A+ is followed by a trial to AY−: (1) On the A+ trial, V_A is incremented by 1.0; (2) on the AY− trial, V_A and V_Y are

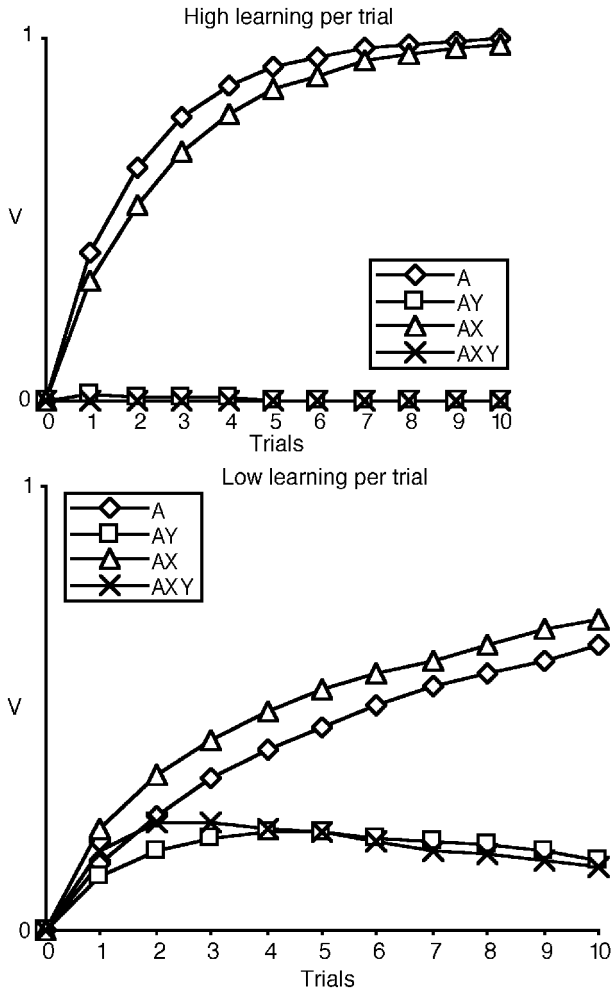

Figure 12. Simulation of a real-time error-correcting learning rule acquiring either the A+, AY− or the AX+, AXY− discrimination, discussed in the text, under conditions of high (upper panel) or low (lower panel) learning per trial. The amount of learning per trial is critical in determining whether the addition of extra common elements to a discrimination will slow acquisition of that discrimination (upper panel) or speed it (lower panel).

decremented by .5 each. Thus, at the end of the second trial, $V_A = .5$ and $V_Y = -.5$, so that $V_A = .5$ and $V_A + V_Y = 0$. Therefore, discrimination between A and AY (a difference in associative strength of 0.5) is better than that between AX and AXY (a difference of 0.33). The addition of unique configural cues does not alter this prediction.

As we noted above, this prediction does not require that the learning per trial be as high as in the example given. Values as low as .3 of asymptotic learning still predict that the addition of a common element will retard discrimination learning, with break-even at around 0.25. Figure 12 shows the results of two simulations, one with learning per trial set at .4 of asymptote and the other set at .1.

This analysis applies equally to several other cases examined by Pearce and his colleagues (e.g., Pearce & Redhead, 1993; Redhead & Pearce, 1995), including, for example, the negative patterning discrimination between A+, BC+, and ABC−, where the straightforward prediction from Rescorla and Wagner (1972) is that the discrimination between BC and ABC should be learned more rapidly than that between A and ABC, whereas a configural analysis predicts (correctly) the opposite outcome. An elemental theory here will, of course, need to postulate unique, configural cues in order to solve this negative patterning problem, but once this is done, our own model, with learning per trial greater than 0.25 of asymptote, correctly predicts better discrimination between A and ABC than between BC and ABC. The reader may feel that, here, we are carefully assuming a value for the amount of learning per trial that allows us to explain Pearce's data, and the reader would be right. It should, however, be possible to test the critical assumption that the outcome of such an experiment will depend on the amount of learning per trial and so distinguish empirically between configural and elemental theories. But note that Pearce is also making important assumptions in order to get his configural model to fit the data. Learning theorists are accustomed to ignoring the effect of context in their calculations, because in most cases, the context will acquire little associative strength if it has zero correlation with the US, and making explicit allowance for the presence of contextual cues has little or no effect on the predictions derivable from an elemental model. But the context cannot be neglected in this way when a configural theory is applied. Including it has the effect of changing many of the configural theory's predictions so that they come into line with those of the Rescorla–Wagner model. In the present instance, if we allow the context sufficient salience, it can become an important influence, so that adding extra common elements to the discrimination predicts facilitated learning by lowering the generalization between the experimental stimuli and the context to the extent that this outweighs the increased generalization between positive and negative components of the discrimination. Thus, in other words, the configural theory will predict faster acquisition of the discrimination when extra common elements are added.

Deletion of Common Elements in Compounds

As we have noted earlier (pp. 192–193) and will argue in more detail below, an elemental theory is not forced to predict that the associative values of two stimuli, A and B, are simply summed to yield the associative value of the AB compound. If elements common to A and B are not represented twice in the compound, there will be incomplete summation. But this also allows us to explain another set of findings reported by Pearce and Wilson (1990).

Consider the following discrimination: A+, AB−, BC+. At asymptote, Pearce and Wilson (1990) derive the prediction for Rescorla–Wagner (without unique cues)

that responding to C should be greater than responding to BC. It is easy to see why. At asymptote,

$$V_A \rightarrow \lambda,$$
$$V_{AB} \rightarrow 0,$$
$$\therefore V_B \rightarrow -\lambda,$$

and

$$V_{BC} \rightarrow \lambda,$$
$$\therefore V_C \rightarrow 2\lambda.$$

Thus, C is a strong excitor, B is an inhibitor, and responding to BC will be less than that to C. Several experiments have established, however, that subjects may respond more to BC after this sort of training than to C (Nakajima, 1997; Pearce & Wilson, 1990).

If, however, we assume distributed representations, so that Stimuli A, B, and C all share elements with one another in a nonadditive fashion, the prediction from elemental theory no longer holds and can even be reversed, in agreement with these experimental data. A simple example serves to demonstrate this. Take the special case in which Stimuli A, B, and C can be broken down into elements a + x + y, b + x + z, and c + y + z, respectively, with x common to A and B, y common to A and C, and z common to B and C. This gives us

$$V_a + V_x + V_y = \lambda \qquad \text{from } V_A \rightarrow \lambda$$
$$V_a + V_b + V_x + V_y + V_z = 0 \qquad \text{from } V_{AB} \rightarrow 0$$
$$V_b + V_c + V_x + V_y + V_z = \lambda \qquad \text{from } V_{BC} \rightarrow \lambda.$$

With some algebra, we can show that

$$V_{BC} = \lambda$$

and

$$V_C = V_c + V_y + V_z = \lambda - (V_b + V_x).$$

Clearly, C may now have a net associative strength less than λ and, hence, less than BC; all that is needed is for V_x to acquire sufficient associative strength to make $V_b + V_x$ positive [since the associative strength for C is $\lambda - (V_b + V_x)$]. As a concrete example of this, consider the case in which $V_a = -\lambda/2$, $V_b = -\lambda/2$, $V_c = \lambda/2$, $V_x = V_y = 3\lambda/4$, and $V_z = -\lambda/2$. With this solution to the equations, C will have a net associative strength of $3\lambda/4$, less than the compound BC, which will have a net associative strength of λ.

Exactly the same elemental solution will yield the same prediction for a slight variation on this problem studied by Nakajima (1997) and Nakajima and Urushihara (1999), in which both rats and pigeons were trained on the discrimination A+, AB−, ABC+. Once again, responding to BC will exceed that to C alone, provided that $V_b + V_x > 0$. But Nakajima and Urushihara took an important further step by varying the modality of the three stimuli, A, B, and C. They showed that the above result held only when all three stimuli were from the same modality or A and B were from the same modality while C was from a different modality. If, on the other hand, A and C were from the same modality while B was from a different modality, then B acted as a conditioned inhibitor when compounded with C—that is to say, animals responded more to C alone than to the BC compound. They showed how it was possible for Pearce's configural theory to explain some of these findings by assuming that two stimuli, A and C, from the same modality shared elements in common—that is, could be represented as a + x and c + x. It is notable, however, that here, too, it was necessary to assume that common elements were not represented twice over in a compound. AC had to be represented as a + c + x.

Although Nakajima and Urushihara (1999) assumed that it was not possible for an elemental theory to predict that the associative value of BC should be greater than that of C alone following such discrimination training, we have shown above that this is not true. Moreover, the same elemental approach can also predict the results of experiments that varied the modality of the three stimuli. We assume, as before, that each of the three stimuli (A, B, and C) are represented by some unique and some shared elements but that stimuli from the same modality share more elements in common, whereas those from different modalities share less. Thus, for the case in which A and C are from the same modality but B is from a different modality, A = a + x + 3y, B = 3b + x + z, C = c + 3y + z, where 3y should be read as $y_1 + y_2 + y_3$, and so on. Then,

$$V_a + V_x + 3V_y = \lambda$$
$$V_a + 3V_b + V_x + 3V_y + V_z = 0$$
$$V_a + 3V_b + V_c + V_x + 3V_y + V_z = \lambda.$$

From this it follows that

$$3V_b + V_z = -\lambda$$
$$V_c = \lambda.$$

Thus, for C we have

$$V_c + 3V_y + V_z,$$

and for BC we have

$$3V_b + V_c + V_x + 3V_y + V_z,$$

a difference of $3V_b + V_x$ between them. If we treat all elements as being of equal salience, we expect V_b to have a value of approximately $-\lambda/4$ and V_x to have a value near $\lambda/5$. Thus, $3V_b + V_x$ will be negative, and responding will be stronger to C than to BC—that is, B will act as an inhibitor, in accordance with the results of Nakajima and Urushihara.

When A and B are from the same modality and C is from a different modality, it follows that the three stimuli will be represented by A = a + 3x + y, B = b + 3x + z, and C = 3c + y + z. Now,

$$V_a + 3V_x + V_y = \lambda$$
$$V_a + V_b + 3V_x + V_y + V_z = 0$$
$$V_a + V_b + 3V_c + 3V_x + V_y + V_z = \lambda.$$

From this, it follows that

$$V_b + V_z = -\lambda$$
$$3V_c = \lambda.$$

Thus, for C we have

$$3V_c + V_y + V_z.$$

And for BC, we have

$$V_b + 3V_c + 3V_x + V_y + V_z.$$

This yields a difference of $V_b + 3V_x$ between them. Since V_x might be expected to have a strength of $\lambda/5$ and V_b a strength of $-\lambda/2$, this means that the difference will be positive and responding to BC can be expected to be greater than that to C.

For all these derivations, the critical assumption has been that the elements common to two component stimuli are not represented twice over in the compound. Nakajima and Urushihara (1999) attempted to use a similar common elements assumption in their application of Pearce's theory to their results, but even then they encountered problems. Pearce's original theory can deal with Nakajima's (1997) results without modification, but, when common elements are added, the prediction that responding to BC should be stronger than that to C alone is reversed. Thus, the theory, when modified to fit the results of one study, loses its ability to give a satisfactory account of the other. It is also worth noting that an appeal to contextual stimuli may not be a solution for Pearce. If they are included in the analysis, the effect is to change the generalization coefficients between the various components of the discrimination and, if the context is sufficiently salient, to change the predictions of this configural model so that they are once again in line with those of the simple Rescorla–Wagner model. This may help with some of the results reported by Nakajima and Urushihara but would invalidate much of Pearce's earlier work.

We also note that the assumption that elements common to A and B are not represented twice over in the AB compound allows a solution to the negative patterning problem without appeal to unique configural cues: The discrimination AX+, BX+, and ABX− is solved if X = 2λ and A = B = $-\lambda$. Rescorla (1972) rejected this solution on the grounds that adding extra common elements to the problem (i.e., moving from an original A+, B+, AB− to AX+, BX+, ABX−) did not facilitate learning as predicted by this analysis. Our model, however, can accommodate the results of both Rescorla and Pearce by assuming high learning per trial, and this allows the retention of the common element explanation of negative patterning. Again, this is not to say that we think that there is no role for unique cues (in the sense postulated by Rescorla) in the solution of negative patterning problems but merely that there is likely to be a role for nonadditive common elements as well in certain circumstances.

Summation

As both Pavlov (1927) and Konorski (1948) observed, conditioning occurs more rapidly, and usually to a higher asymptote, the more intense or salient the CS (see also Kamin, 1965). And, as we have already seen in experiments on transfer along a continuum, discrimination between two widely separated stimuli is easier than that between two stimuli lying closer together on the same continuum. Conditioning also often proceeds more rapidly to a compound CS (e.g., one with a visual and an auditory component) than to either component alone (e.g., Kamin, 1969; Mackintosh, 1976), and a discrimination with two sets of relevant cues (e.g., between a pair of stimuli differing in both brightness and orientation, such as a black horizontal rectangle vs. a white vertical rectangle) is learned more rapidly than the discrimination between either component alone (Miles & Jenkins, 1973; Sutherland & Mackintosh, 1971). Finally, numerous studies of both Pavlovian conditioning and discrimination learning have found that if animals receive separate conditioning trials to each of two component CSs (a light and a noise) or learn two concurrent discriminations (e.g., between black and white squares and between horizontal and vertical gray rectangles), then, provided performance is not asymptotic at the end of training, test trials with both component CSs presented in compound or both discriminative cues presented together often yield performance superior to that attained during training with the component stimuli. This is the phenomenon of summation in conditioning (Kehoe & Gormezano, 1980; Konorski, 1948; see Myers, Vogel, Shin, & Wagner, 2001, for a recent demonstration) or additivity of cues in discrimination learning (McGonigle, 1967; Sutherland & Mackintosh, 1971).

Elemental models, such as ours or Rescorla and Wagner's (1972), have no problem predicting all these findings. The intensity or salience of a CS or discriminative stimulus will be positively related to the rate of learning. Conditioning to a compound CS or discrimination between compound discriminative stimuli will normally proceed more rapidly than it will to the individual components, because (except in the case in which one component is so much more salient than the other that it completely overshadows it) each component will acquire some associative value and these values will summate to produce not only faster acquisition to the compound, but also summation after training on each component alone.

Pearce (1987) allows that stimuli may differ in salience and, thus, has no difficulty predicting the effect of salience on conditioning or of the distance between S+ and S− on speed of discrimination learning. By assuming that an AB compound will be more salient than either component alone, he can also predict that conditioning will normally proceed more rapidly to a compound CS than to either of its components and that the discrimination between two compound stimuli will be learned more rapidly than that between either of the components. But straightforward application of the theory predicts no summation to a compound after conditioning to its components and, by the same token, no additivity of cues after discrimination training to two separate sets of discriminative stimuli. If animals receive

conditioning trials to two component CSs, A and B, and are then tested on AB, the compound will receive only generalized associative strength from A and B (the AB configuration has never itself been conditioned), and that generalization will be determined by the similarity of AB to A and B. Assuming that A and B are equally salient, the compound shares half its elements with each component and, therefore, receives half the associative strength of each. Thus, the total value of AB will be exactly the same as that of either A or B alone.

Although it might seem, therefore, that experiments on summation and additivity of cues lend support to elemental, rather than configural, theories, that conclusion requires immediate qualification. The fact is that summation is not always observed: Failure to find evidence of summation has been reported in a number of studies (e.g., Aydin & Pearce, 1995; Kehoe, Horne, Horne, & Macrae, 1994; Pearce, George, Redhead, Aydin, & Wynne, 1999; Rescorla, 1997, 1999; Rescorla & Coldwell, 1995). Thus, the first question at issue is whether it is possible to specify the conditions that yield summation and those that do not and whether any particular theory is able to offer a principled explanation of this variation in experimental outcome.

Several studies have made it clear that summation is more reliably observed when A and B are stimuli from different modalities than when both are from the same modality (e.g., Aydin & Pearce, 1997; Kehoe et al., 1994; Rescorla & Coldwell, 1995). Indeed, the experimental paradigm that has been least successful in yielding evidence of summation has been pigeon autoshaping, when A and B are localized visual stimuli projected onto a response panel. As Myers et al. (2001) have argued, elemental theories have a fairly natural way of explaining why summation is less likely to occur when A and B are stimuli from the same modality. If A and B are similar, they share elements in common—that is to say, they can be represented as AX and BX. As we have already argued, when two similar stimuli are presented in compound, their common elements are not represented twice over; in other words, the AB compound will be represented as ABX, rather than as ABXX. Thus, the associative strength of the AB compound will not be equal to the sum of the associative strengths of A and B. Moreover, the more similar A and B are and the higher the proportion of X to A and B elements, the less summation will be predicted. In the terminology of Wagner and Brandon's (2001) elemental analysis of elemental and configural theories, this assumption amounts to saying that half the elements common to two component stimuli are inhibited when the components are presented in compound. As Myers et al. (2001) have pointed out, on that analysis, the presence or absence of summation is dependent on the proportion of elements that are inhibited when two components are compounded. If a sufficient proportion of the unique elements of each component are also inhibited in the compound, no summation at all will be predicted.

In contrast, Pearce's (1987, 1994) configural theory incorrectly predicts that summation between two similar stimuli, AX and BX, will be *greater* than that between two dissimilar stimuli, A and B. As we have already seen, assuming equal salience of A and B, Pearce predicted that the generalized value of the AB compound would be equal to the value of either A or B. But by assuming equal salience for A, B, and X, Pearce's model predicts that, at asymptote, $V_{AX} = V_{BX} = 0.8\lambda$ but that the generalization from each to ABX will be .67, so $V_{ABX} = 1.07\lambda$. However, this appeal to common elements shared by stimuli in the same modality does allow the prediction of summation from Pearce's theory in other cases. For example, Pearce, George, and Aydin (2002) replicated an earlier study by Rescorla (1997) in which rats were trained to respond to two component stimuli, A and B, and concurrently to a third, compound stimulus, CD. Since A and B were stimuli from different modalities, the straightforward prediction from Pearce's theory is that $V_{AB} = V_{CD} = \lambda$. But Pearce et al. argued that, since both AB and CD consisted of one auditory and one visual stimulus, there would have been some generalization between the two compounds and this generalization from CD to AB could have been sufficient to generate a higher level of responding to AB than to CD. Consistent with this analysis, experiments that reduced the associative value of the CD compound resulted in the animals responding no more to AB than to either A or B alone.

A second way for Pearce's configural theory to predict summation is to appeal to contextual cues. Just as any other elements common to both A and B will increase generalization from A and B to the AB compound and so allow the prediction of some summation, so contextual cues, if sufficiently salient, will also increase generalization from A and B to the compound. Let X represent the contextual cues, then reinforcement of A and B is represented as AX+, BX+, X−; assuming equal salience of A, B, and the contextual cues, at asymptote $V_{AX} = V_{BX} = 1.33\lambda$ and $V_X = -1.33\lambda$, with the inhibition generalizing from X to AX and BX, keeping their total value at λ. But since the generalization from both AX and BX to ABX is 0.67 and from X to ABX is only 0.33, $V_{ABX} = 1.33\lambda$. Although Pearce, George, Redhead, Aydin, and Wynne (1999) have presented some evidence from a study of pigeon autoshaping that lends some support to this analysis, the appeal to contextual stimuli is a double-edged one, because it can wreak havoc with some of the other predictions from Pearce's configural theory. Many of the predictions derived from Pearce's theory that distinguish it from those of the Rescorla–Wagner model of Pavlovian conditioning are changed so as to agree with Rescorla–Wagner if contextual cues are given sufficient salience in the analysis.

There remains, as Pearce et al. (2002) have acknowledged, other evidence of summation that does not yield readily to a configural analysis (e.g., Ganesan & Pearce, 1988; Rescorla, 1997, Experiments 2 and 4). We are inclined to believe, however, that the problem is wider than

Table 1
Experimental Design

Group	Conditioning Stimuli	Test Stimuli
Experimental	AC+ | AD+ | AE+ | AF+ | A+ BC+ | BD+ | BE+ | BF+ | B+	AB
Control	A+ B+	AB

this. With some effort, the configural account can be made to predict modest amounts of summation. But there are numerous experiments that suggest that, early in training or when other steps are taken to ensure relatively modest levels of responding to A and B, the summation observed to the AB compound can be very substantial indeed (e.g., Kehoe, 1986; Kehoe et al., 1994; Konorski, 1948; Sutherland & Mackintosh, 1971). Indeed, in rabbit eyelid conditioning and simultaneous visual discrimination learning in rats, responding to the compound is well predicted by the formula for combining two independent probabilities:

$$P(\mathrm{AB}) = P(\mathrm{A}) + P(\mathrm{B}) - [P(\mathrm{A}) \times P(\mathrm{B})]. \qquad (2)$$

As Kehoe argued, this is exactly the outcome predicted by a strictly elemental theory.

Our account of summation, like that of Wagner and Brandon (2001), is sufficiently flexible that it can explain both the occurrence of summation and its absence. Since the available data suggest that summation sometimes occurs but sometimes fails to occur, this might seem to be a virtue. A critic could reasonably respond, however, that a theory capable of explaining any possible experimental outcome is hardly worth having. To be of any value, the theory must be able to specify when summation will and will not occur. Myers et al. (2001) agreed with Rescorla and Coldwell (1995) in laying emphasis on the nature of the stimuli. Put loosely, summation will fail to occur to an AB compound when there is a perceptual interaction between A and B. It is under these circumstances that there will be deletion not only of some of the elements common to A and B when the compound is presented, but also of some of their unique elements. Consistent with this suggestion, the preparation that has most consistently failed to yield evidence of summation is pigeon autoshaping to localized visual stimuli presented on a response panel. As Rescorla and Coldwell showed, pigeons are more likely to show summation when at least one of the two components is a diffuse auditory or visual stimulus. We follow them in this, but the problem is to specify what is meant by *perceptual interaction*. Myers et al. refer to the distinction drawn some years ago by Garner (1974) between integral and separable stimuli, the former showing perceptual interactions, the latter not. It remains to be seen whether this line of inquiry can be profitably pursued.

Our own theory suggests at least one further possibility. We have distinguished between *core* and *peripheral* units activated by the presentation of a stimulus. The core units are those activated regardless of the context in which the stimulus occurs and of any other stimuli presented in conjunction with it. The peripheral units are precisely those whose level of activation is markedly affected by contextual and other stimuli. Summation between A and B will occur to the extent that conditioning to them results in the acquisition of associative strength by the core units, rather than by any peripheral units, since the activation of peripheral units will change when the AB compound is presented. It should be possible to assess this prediction empirically. Consider the experimental design illustrated in Table 1. The control group provides a standard test of summation. Following separate conditioning to A and B, they are tested for responding to the AB compound. Depending on the nature of the stimuli, our elemental analysis will predict either substantial or only marginal summation, but for the purposes of this experiment, it will be important to choose stimuli that yield no more than modest summation. The experimental group receives exactly the same total number of reinforced trials to A and B as the control group, but for the initial phase of this training, A and B are presented in compound with a series of other stimuli (C, D, E, and F). According to our analysis, each time A or B is presented in compound with a new second stimulus, there will be a change in the activation of their peripheral units. Only their core units will maintain a stable level of activation throughout this phase of the experiment, and they will consequently acquire the major part of the associative value available for conditioning. Therefore, we predict that the experimental group will show much more summation than the control group, because the units carrying substantial associative strength for the subjects in the experimental group will be exactly those less affected by compounding A and B on test.

In more general theoretical terms, the critical difference between our analysis of the experimental group's treatment and that provided by Pearce's theory is that Pearce's model proliferates configural representations that then generalize to give the net associative strength for A or B, whereas our model predicts that only a subset of the units representing A or B will carry the necessary associative strength. Pearce's theory must inevitably involve more generalization from a larger set of units as learning proceeds, whereas much of the time, our elemental approach predicts that the number of units crucial to learning will decrease as a consequence of learning. This difference between the extensive and the intensive is fundamental and underpins the difference between our use of units sensitive to configurations of perceptual elements and the status of Pearce's model as containing a genuinely configural level of representation.

CONCLUSION

Our general argument has been that a fully worked out elemental theory of the representation of stimuli is powerful enough to explain not only the major empirical phenomena of generalization, widely seen as amenable to el-

emental analysis, but also more particular results, such as those pertaining to peak shift and transfer along a continuum. Such a theory can also be flexible enough to explain a variety of phenomena thought to require a configural analysis—not only the solution of various configural and conditional discriminations, but also the role of similarity in discrimination learning and summation.

We do not pretend that, as it stands, it will be able to adequately explain every detail of human or even animal performance on these tasks. There are phenomena outside the scope of our model; attentional effects (e.g., learned irrelevance) in animals constitute one example. But these limitations notwithstanding, we believe that the model described here and in the earlier companion paper (McLaren & Mackintosh, 2000) is an account of elemental representation that will have wide application in the domain of associative learning.

REFERENCES

Atkinson, R. C., & Estes, W. K. (1963). Stimulus sampling theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 2, pp. 121-268). New York: Wiley.

Aydin, A., & Pearce, J. M. (1995). Summation in autoshaping with short- and long-duration stimuli. *Quarterly Journal of Experimental Psychology*, **48B**, 215-234.

Bennett, C. H., Wills, S. J., Wells, J. O., & Mackintosh, N. J. (1994). Reduced generalization following preexposure: Latent inhibition of common elements or a difference in familiarity? *Journal of Experimental Psychology: Animal Behavior Processes*, **20**, 232-239.

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

Bush, R. R., & Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review*, **58**, 313-323.

Estes, W. K. (1959). The statistical approach to learning theory. In S. Koch (Ed.), *Psychology: A study of a science* (Vol. 2, pp. 380-491). New York: McGraw-Hill.

Ganesan, R., & Pearce, J. M. (1988). Interactions between conditioned stimuli for food and water in the rat. *Quarterly Journal of Experimental Psychology*, **40B**, 229-241.

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.

George, D. N., Ward-Robinson, J., & Pearce, J. M. (2001). Discrimination of structure: I. Implications for connectionist theories of discrimination learning. *Journal of Experimental Psychology: Animal Behavior Processes*, **27**, 206-218.

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

Gulliksen, H., & Wolfle, H. L. (1938). A theory of learning and transfer: I. *Psychometrika*, **3**, 127-149.

Hall, G. (1980). Exposure learning in animals. *Psychological Bulletin*, **88**, 535-550.

Hall, G. (1991). *Perceptual and associative learning*. Oxford: Oxford University Press, Clarendon Press.

Hanson, H. M. (1959). Effects of discrimination training on stimulus generalization. *Journal of Experimental Psychology*, **58**, 321-344.

Hearst, E. (1968). Discrimination learning as the summation of excitation and inhibition. *Science*, **162**, 1303-1306.

Honig, W. K., & Urcuioli, P. J. (1981). The legacy of Guttman and Kalish (1956): Twenty-five years of stimulus generalization research. *Journal of the Experimental Analysis of Behavior*, **36**, 405-445.

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

Hull, C. L. (1952). *A behavior system*. New Haven, CT: Yale University Press.

James, W. (1890). *Principles of psychology*. New York: Holt.

Jenkins, H. M., & Harrison, R. H. (1960). Effect of discrimination training on auditory generalization. *Journal of Experimental Psychology*, **59**, 246-253.

Jenkins, H. M., & Harrison, R. H. (1962). Generalization gradients of inhibition following auditory discrimination learning. *Journal of the Experimental Analysis of Behavior*, **5**, 435-441.

Jones, F., & McLaren, I. P. L. (1999). Rules and associations. In *Proceedings of the Twenty-First Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Kamin, L. J. (1965). Temporal and intensity characteristics of the conditioned stimulus. In W. F. Prokasy (Ed.), *Classical conditioning: A symposium* (pp. 118-147). New York: Appleton-Century-Crofts.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior* (pp. 279-296), New York: Appleton-Century Crofts.

Kehoe, E. J. (1986). Summation and configuration in conditioning of the rabbit's nictitating membrane response to compound stimuli. *Journal of Experimental Psychology: Animal Behavior Processes,* **12**, 186-195.

Kehoe, E. J., & Gormezano, I. (1980). Configuration and combination laws in conditioning with compound stimuli. *Psychological Bulletin*, **87**, 351-378.

Kehoe, E. J., Horne, A. J., Horne, P. S., & Macrae, M. (1994). Summation and configuration between and within sensory modalities in classical conditioning of the rabbit. *Animal Learning & Behavior*, **22**, 19-26.

Klein, M., & Rilling, M. (1974). Generalization of free-operant avoidance behavior in pigeons. *Journal of the Experimental Analysis of Behavior*, **21**, 75-88.

Konorski, J. (1948). *Conditioned reflexes and neuron organisation*. Cambridge: Cambridge University Press.

Lawrence, D. H. (1952). The transfer of a discrimination along a continuum. *Journal of Comparative & Physiological Psychology*, **45**, 511-516.

Lawrence, D. H. (1955). The applicability of generalization gradients to the transfer of a discrimination. *Journal of General Psychology*, **52**, 37-48.

Logan, F. A. (1966). Transfer of discrimination. *Journal of Experimental Psychology*, **71**, 616-618.

Mackintosh, N. J. (1974). *The psychology of animal learning*. London: Academic Press.

Mackintosh, N. J. (1976). Overshadowing and stimulus intensity. *Animal Learning & Behavior*, **4**, 186-192.

Mackintosh, N. J., Kaye, H., & Bennett, C. H. (1991). Perceptual learning in flavour aversion conditioning. *Quarterly Journal of Experimental Psychology*, **43B**, 297-322.

Mackintosh, N. J., & Little, L. (1970). An analysis of transfer along a continuum. *Canadian Journal of Psychology*, **24**, 362-369.

Marsh, G. (1972). Prediction of the peak shift in pigeons from gradients of excitation and inhibition. *Journal of Comparative Physiological Psychology*, **81**, 262-266.

McGonigle, B. (1967). Stimulus additivity and dominance in visual discrimination performance by rats. *Journal of Comparative & Physiological Psychology*, **64**, 110-113.

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). Oxford: Oxford University Press, Clarendon Press.

McLaren, I. P. L., & Mackintosh, N. J. (2000). An elemental model of associative learning: I. Latent inhibition and perceptual learning. *Animal Learning & Behavior*, **28**, 211-246.

McLaren, I. P. L., & Suret, M. (2000). Transfer along a continuum: Differentiation or association? In *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Medin, D. L. (1975). A theory of context in discrimination learning. In G. H. Bower (Ed.). *The psychology of learning and motivation* (Vol. 9, pp. 263-314). New York: Academic Press.

Miles, C. G., & Jenkins, H. M. (1973). Overshadowing in operant conditioning as a function of discriminability. *Learning & Motivation*, **4**, 11-27.

Myers, K. M., Vogel, E. H., Shin, J., & Wagner, A. R. (2001). A comparison of the Rescorla–Wagner and Pearce models in a negative patterning and a summation problem. *Animal Learning & Behavior*, **29**, 36-45.

Nakajima, S. (1997). Failure of inhibition by B over C after A+, AB−, ABC+ training. *Journal of Experimental Psychology: Animal Behavior Processes*, **23**, 482-490.

Nakajima, S., & Urushihara, K. (1999). Inhibition and facilitation by B over C after A+, AB−, and ABC+ training with multimodality stimulus combinations. *Journal of Experimental Psychology: Animal Behavior Processes*, **25**, 68-81.

Oakeshott, S. M., & Mackintosh, N. J. (2002). *Peak shift and absence of peak shift along an artificial dimension.* Manuscript in preparation.

Pavlov, I. P. (1927). *Conditioned reflexes* (G. V. Anrep, Trans.). London: Oxford University Press.

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

Pearce, J. M. (1994). Similarity and discrimination: A selective review and a connectionist model. *Psychological Review*, **101**, 587-607.

Pearce, J. M., George, D. N., & Aydin, A. (2002). Summation: Further assessment of a configural theory. *Quarterly Journal of Experimental Psychology*, **55B**, 61-73.

Pearce, J. M., George, D. N., Redhead, E. S., Aydin, A., & Wynne, C. (1999). The influence of background stimuli on summation during autoshaping. *Quarterly Journal of Experimental Psychology*, **52B**, 53-74.

Pearce, J. M., & Redhead, E. S. (1993). The influence of an irrelevant stimulus on two discriminations. *Journal of Experimental Psychology: Animal Behavior Processes*, **19**, 180-190.

Pearce, J. M., & Wilson, P. N. (1990). Configural associations in discrimination learning. *Journal of Experimental Psychology: Animal Behavior Processes*, **16**, 250-261.

Redhead, E. S., & Pearce, J. M. (1995). Similarity and discrimination learning. *Quarterly Journal of Experimental Psychology*, **48B**, 46-66.

Rescorla, R. A. (1972). "Configural" conditioning in discrete-trial bar pressing. *Journal of Comparative & Physiological Psychology*, **79**, 307-317.

Rescorla, R. A. (1973). Evidence for "unique stimulus" account of configural conditioning. *Journal of Comparative & Physiological Psychology*, **85**, 331-338.

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

Rescorla, R. A. (1997). Summation: Assessment of a configural theory. *Animal Learning & Behavior*, **25**, 200-209.

Rescorla, R. A. (1999). Summation and overexpectation with qualitatively different outcomes. *Animal Learning & Behavior*, **27**, 50-62.

Rescorla, R. A., & Coldwell, S. E. (1995). Summation in autoshaping. *Animal Learning & Behavior*, **23**, 314-326.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.

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.

Saavedra, M. A. (1975). Pavlovian compound conditioning in the rabbit. *Learning & Motivation*, **6**, 314-326.

Saksida, L. M. (1999). Effects of similarity and experience on discrimination learning: A nonassociative connectionist model of perceptual learning. *Journal of Experimental Psychology: Animal Behavior Processes*, **25**, 308-323.

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

Spence, K. W. (1936). The nature of discrimination learning in animals. *Psychological Review*, **43**, 427-449.

Spence, K. W. (1937). The differential response in animals to stimuli varying within a single dimension. *Psychological Review*, **44**, 430-444.

Suret, M., & McLaren, I. P. L. (2002). Associative modelling of human learning on an artificial dimension. In *Proceedings of the WCCI*. Hawaii.

Sutherland, N. S., & Mackintosh, N. J. (1971). *Mechanisms of animal discrimination learning*. New York: Academic Press.

Thompson, R. F. (1965). The neural basis of stimulus generalization. In D. I. Mostofsky (Ed.), *Stimulus generalization* (pp. 154-178). Stanford: Stanford University Press.

Wagner, A. R. (1981). SOP: A model of automatic memory processing in animal behavior. In N. E. Spear & R. R. Miller (Eds.), *Information processing in animals: Memory mechanisms* (pp. 95-128). Hillsdale, NJ: Erlbaum.

Wagner, A. R., & Brandon, S. E. (2001). A componential theory of Pavlovian conditioning. In R. R. Mowrer & S. B. Klein (Eds.), *Handbook of contemporary learning theories* (pp. 23-64). Mahwah, NJ: Erlbaum.

Wagner, A. R., & Rescorla, R. A. (1972). Inhibition in Pavlovian conditioning: Application of a theory. In R. A. Boakes & M. S. Halliday (Eds.), *Inhibition and learning* (pp. 301-336). London: Academic Press.

Wills, S. J., & Mackintosh, N. J. (1998). Peak shift on an artificial dimension. *Quarterly Journal of Experimental Psychology*, **51B**, 1-31.

Woodbury, C. B. (1943). The learning of stimulus patterns by dogs. *Journal of Comparative Psychology*, **35**, 29-40.

(Manuscript received October 1, 2001;
revision accepted for publication April 17, 2002.)