Published in *Psychological Review*, vol 113, 584-605 http://www.apa.org/pubs/journals/xan/index.aspx

Elemental Representations of Stimuli in Associative Learning

Justin A. Harris
The University of Sydney

Abstract

This paper reviews evidence and theories concerning the nature of stimulus representations in Pavlovian conditioning. It focuses on the elemental approach developed in Stimulus Sampling Theory (Atkinson & Estes, 1963; Bush & Mosteller, 1951b) and extended by McLaren and Mackintosh (2000; 2002), and contrasts this with models that that invoke notions of configural representations that uniquely code for different patterns of stimulus inputs (e.g., Pearce, 1987, 1994; Rescorla & Wagner, 1972; Wagner & Brandon, 2001). The paper then presents a new elemental model that emphasizes interactions between stimulus elements. This model is shown to explain a range of behavioral findings, including those (e.g., negative patterning and biconditional discriminations) traditionally thought beyond the explanatory capabilities of elemental models. Moreover, the model offers a ready explanation for recent findings reported by Rescorla (2000; 2001; 2002b) concerning the way that stimuli with different conditioning histories acquire associative strength when conditioned in compound.

Address correspondence to:

Dr Justin Harris
School of Psychology
University of Sydney
Sydney 2006
Australia.

Email: justinh@psych.usyd.edu.au
Fax: (++61 2) 9351 2603 or 9036 5223
Tel: (++61 2) 9351 2864

Key words:
Pavlovian conditioning; configural; stimulus sampling theory; discrimination learning.

Author note:

The author wishes to thank Fred Westbrook, Bob Boakes, Peter Lovibond, and Chris Mitchell for fruitful discussions and comments, as well as three anonymous reviewers whose instructive and constructive comments greatly improved the quality and depth of the manuscript.

Pavlovian conditioning has come to be viewed as the cardinal example of associative learning – the process by which an organism represents the correlations between the events it experiences. Efforts to describe this process have typically approached the task by breaking it into two separable but related problems. The first concerns the nature of associations themselves: what is the content of associations, what are the conditions that promote their formation, and how are they expressed in the organism's behavior? The second problem concerns the nature of the stimulus representations between which associations form, since any understanding of associative learning will depend on an appropriate description of these representations. It is this second problem that forms the focus of the present paper.

Empirical evidence about the nature of stimulus representations is largely derived from experiments examining how the conditioned response (CR) to one stimulus, or a compound of two or more stimuli, generalizes to another stimulus or stimulus compound (Kehoe & Gormezano, 1980). Based on such evidence, different theoretical positions have been put forward to describe the mechanisms underlying stimulus representation, but essentially all operate within one (or both) of two frameworks. One framework, most clearly exemplified by Stimulus Sampling Theory (Atkinson & Estes, 1963; Bush & Mosteller, 1951b; Estes, 1950), treats stimulus patterns as comprised of elemental units each of which enters into the associative structure. The other framework treats stimulus patterns as distinct configurations, such that associations operate on the configuration as a whole. An intermediate approach combines aspects of both frameworks. For example, the Rescorla-Wagner model adopts an elemental framework to explain the interaction between stimulus representations and associations, but incorporates the notion that compounds of two or more stimuli are represented by a configural element unique to the compound in addition to the individual elements that comprise each of the stimuli (Rescorla & Wagner, 1972; Wagner & Rescorla, 1972).

The objective of the present paper is to describe how stimulus representation can be understood within a strictly elemental framework, and how to reconcile evidence that has traditionally been held to contradict the elemental view. The paper begins with a brief description of Stimulus Sampling Theory and the Rescorla-Wagner model that incorporated a similar elemental approach to stimulus representation. I then review findings that have been held up as evidence against the elemental approach, and as support for the alternative configural view. The main focus of the paper is the presentation of a new elemental model in which stimuli are represented as an array of elemental units that correspond to different stimulus features. The model describes mechanisms by which individual elements interact to influence each other's activation and entry to a limited-capacity attention buffer. These interactions affect both conditioning to the individual elements and responding provoked by those elements.

Stimulus Sampling Theory

Stimulus Sampling Theory, in its various formulations, constituted the first comprehensive treatment of stimulus representations in Pavlovian conditioning (Atkinson & Estes, 1963; Bush & Mosteller, 1951b; Estes, 1950). It remains the most successful account of what is the most fundamental problem for any model of stimulus representation – discrimination versus generalization between stimuli. Responding to a conditioned stimulus (CS) often generalizes to other stimuli, and the degree of this generalization follows what might be described as the similarity of these stimuli to the CS. Stimulus Sampling Theory formalized the notion that stimuli are represented by arrays of elemental features, each of which can independently enter into an association with the unconditioned stimulus (US), and that generalization from one stimulus to another is attributed to overlap in the population of elements that comprise the two stimuli (Bush & Mosteller, 1951b; Estes, 1950). That is, a new stimulus is expected to elicit CRs to the extent that it contains elements that are also present in the CS and so will have undergone conditioning. The generalization gradient is a direct function of the number of elements common to the CS and test stimulus. The absence of an alternative coherent account for stimulus generalization has ensured that this basic mechanism is retained even within theories that invoke

configural representations of stimuli (e.g., Pearce, 1987, 1994; Wagner & Brandon, 2001).

In its original form, Stimulus Sampling Theory had a number of serious limitations, stemming largely from its description of the associative changes provoked by the presence or absence of reinforcement. Its simple linear-operator rule assumed that elements gain or lose associative strength as a function of the discrepancy between their existing state (conditioned or non-conditioned) and that afforded by the current reinforcement (Bush & Mosteller, 1951b; Estes, 1950). In other words, conditioning of any element was independent of the associative status of other elements. Not surprisingly, the theory had no means of anticipating interactions between CSs, such as in Kamin's (1968) demonstration that conditioning to a stimulus could be blocked if the stimulus were reinforced in compound with a previously conditioned stimulus. Of particular relevance to the current review, the theory was unable to offer a satisfactory description of how animals could show errorless performance after training on a simple discrimination between stimuli with overlapping representations (Bush & Mosteller, 1951a; Pearce, 1994). Not surprisingly, the model was also unable to explain how animals solve the more complex tasks, such as feature negative discriminations or negative patterning, that have played a key role in shaping recent theories of stimulus representation. As discussed in the next section, many of these limitations can be overcome if the theory is combined with a more sophisticated rule for describing changes in the strength of CS-US associations, such as that proposed by Rescorla and Wagner (1972).

Elemental representations in the Resorla-Wagner model

A key feature of the associative learning model proposed by Rescorla and Wagner (1972) is its treatment of CS representations as distinct units that form independent associations with the US. The model proposes an error-correction rule (Equation 1) to describe how CSs gain or lose associative strength (V) across the course of conditioning.

$$\Delta V_x = \alpha_x \times \beta \times (\lambda - \Sigma V_i) \qquad (1)$$

On a trial-by-trial basis, the strength of the association between a given CS, x, and the US changes in proportion to the discrepancy between the existing associative strength of all *n* CSs present on that trial (ΣV) and the maximum associative strength supported by the US (λ): Small discrepancies provoke small changes in V whereas large discrepancies provoke larger changes in V; a positive discrepancy provokes excitatory learning, a negative discrepancy provokes inhibitory learning or extinction. The actual change in strength (ΔV) is a product of this discrepancy and parameters related to the salience of the CS in question (α) and US (β). The significant advance of the Rescorla-Wagner model was to recognize that, on any conditioning trial, a single discrepancy is calculated based on all CSs present. If two CSs, each with an existing association with the US, are presented together, then further learning will be limited by their summed associative strengths, and thus each CS effectively reduces conditioning to the other. By constraining learning to the computation of this common error term, the model provides an immediate explanation for Kamin's (1968) blocking effect. It also provides the means by which a stimulus could acquire net inhibitory strength, a property that has proved crucial to explaining performance in many discrimination tasks. The "additivity rule" on which the common error term is based assumes that the representations of each CS are separate and form independent associations with the US. In this sense, the Rescorla-Wagner model adopted a basic elemental approach. But because the model was not concerned with providing a detailed description of stimulus representations, it treated each stimulus as a single elemental unit (rather than an array of component elements as assumed in Stimulus Sampling Theory).

The basic elemental approach inherent in the Rescorla-Wagner model means that its associative rule can be readily combined with more detailed models of stimulus representation. Indeed, soon after the Rescorla-Wagner model was proposed, Blough (1975) and Rescorla (1976) showed how its associative learning rule could be successfully combined with Stimulus Sampling Theory to generate a number of novel, empirically verified, predictions. In addition to the benefits

identified by Blough and Rescorla, the combination of these models can be seen to provide a comprehensive explanation for overshadowing – the decrease in conditioning to each of two CSs that are conditioned in compound compared with that produced when the CSs are conditioned individually (Mackintosh, 1976; Pavlov, 1927). The Rescorla-Wagner model explains overshadowing as due to the fact that conditioning to each of two CSs in compound is limited by the associative strength already acquired by the other CS. However, this mechanism cannot explain instances of overshadowing between CSs after a single conditioning trial (James & Wagner, 1980; Mackintosh & Reese, 1979). In this regard it is fortunate that Stimulus Sampling Theory offers a complementary mechanism for overshadowing, one that explains competition between CSs early in the course of conditioning, but cannot explain the persistence of overshadowing across extended conditioning. Stimulus Sampling Theory can explain overshadowing by assuming that, on a trial when two CSs are presented in compound, the resultant increase in the total number of elements present in the conditioning situation decreases the probability of sampling elements of each CS (e.g., Bush & Mosteller, 1951b; Estes, 1950), thus reducing the opportunity for elements of either CS to undergo conditioning.

Two problems for elemental theories

The core assumption of any elemental theory, including Stimulus Sampling Theory and the Rescorla-Wagner model, is that stimulus elements become independently associated with the US. But the soundness of this assumption is seriously questioned by demonstrations that animals can learn certain conditional discriminations between stimulus compounds that cannot be solved by a simple elemental process (Spence, 1952). Two examples are negative patterning and biconditional discrimination. In negative patterning, two CSs (A and B) are presented on separate trials and each is followed by the US (+). Intermixed among these A+ and B+ trials are trials in which the two stimuli are presented simultaneously but not followed by the US (AB−). A simple elemental view predicts that the associative strengths of A and B will increase on A+ and B+ trials, and will decrease on AB− trials. However, responding should always be greater on the AB− trials because the combined associative strengths of A and B will provoke more responding than will that elicited by either stimulus alone. In view of this prediction, it is important that, in a variety of conditioning paradigms, animals have been shown to master negative patterning discriminations, albeit with considerable difficulty: They learn to respond more on A+ and B+ trials than on AB− trials (e.g., Pavlov, 1927; Rescorla, 1972, 1973; Whitlow & Wagner, 1972).

Biconditional discriminations present an even more complex task. Four distinctive stimuli (A, B, C, and D) are presented in four different pairwise combinations, two of which are reinforced (AB+ and CD+) and the other two are not reinforced (AC− and BD−). Thus each of the four CSs is reinforced when presented in one compound and not reinforced in another compound. Therefore, all stimuli have equivalent reinforcement history and so provide no differential information to cue the animal to respond or not respond. In other words, like negative patterning, biconditional discriminations are not solvable by a simple elemental mechanism. Nonetheless, animals can solve such discriminations, learning to respond more on AB+ and CD+ trials than AC− and BD− trials (Rescorla, Grau, & Durlach, 1985; Saavedra, 1975).

The configural solution to conditional discriminations

In light of demonstrations that animals can solve negative patterning and biconditional discriminations, Wagner and Rescorla (1972) adopted a notion put forward by Spence (1952) that stimulus compounds are represented by their components and an additional "configural element" that represents the conjunction of those stimuli. These configural representations function like other elements, in that they enter into associations with the US in the same way that the individual stimulus elements do. This principle is illustrated in Figure 1.

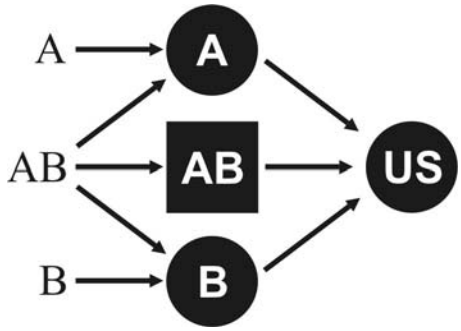

Figure 1. Elemental and configural representations in the Rescorla-Wagner model. Individual stimuli (A and B) activate representations (black circles) that become associated with the US. A compound stimulus (AB) activates these same representations but additionally activates a configural representation (black square) that specifically codes for the conjunction of A and B, and forms an independent association with the US.

The inclusion of this configural representation allows the Rescorla-Wagner model to explain negative patterning and biconditional discriminations. In the case of negative patterning, the separate A and B elements acquire excitatory associative strength with the US, while the AB configural element acquires a strong inhibitory association that opposes the excitatory associations simultaneously activated by the A and B elements. Similarly, a biconditional discrimination is solved by the acquisition of strong inhibitory associations from the configural elements of the non-reinforced compounds (AC and BD in the above example) and excitatory associations from the configural elements of the reinforced compounds (AB and CD).

The addition of a configural element to the representation of a stimulus compound enabled the Rescorla-Wagner model to explain how animals can learn to withhold responses to a non-reinforced compound whose component CSs are excitatory. However, in the absence of such explicit training, the model predicts the summation of responding when two or more CSs are presented in compound. While there are many demonstrations that animals do respond to the compound of two CSs more than to each CS individually (e.g., Kehoe, 1982, 1986; Rescorla, 1997), there are also many reported failures to observe summation in Pavlovian conditioning paradigms. Many of these failures have arisen in autoshaping experiments with pigeons (e.g., Aydin & Pearce, 1995, 1997; Rescorla & Coldwell, 1995), but both successes and failures to observe summation have been reported in other paradigms, such as the conditioned nictitating membrane response in rabbits (Kehoe, Horne, Horne, & Macrae, 1994) and conditioned magazine approach with rats (Pearce, George, & Aydin, 2002; Rescorla, 1997). Such mixed evidence is troubling for the Rescorla-Wagner model because of its commitment to predicting summation.

The replaced elements theory

Brandon and Wagner (1998; Wagner & Brandon, 2001) have recently presented a more elaborate elemental model of stimulus representation that is designed to deal with many of the difficulties that face the Rescorla-Wagner model. Like the approach originally adopted by Wagner and Rescorla (1972), this theory assumes that new configural elements are activated when stimuli are presented in compound, but it additionally proposes the inhibition of elements otherwise activated when the stimuli are presented in isolation. Thus, some elements activated by the individual stimuli are "replaced" by the configural elements activated by the stimulus compound. In its original version, the model assumed a specific pairwise replacement that was different for different compounds (Brandon & Wagner, 1998; Wagner & Brandon, 2001). A simpler and less restrictive replacement process has since been described by Wagner (2003). In this scheme, a sample of elements representing stimulus A are replaced when A is compounded with B, and a statistically independent (and thus potentially overlapping) sample of A elements are replaced when A is compounded with C. The replacement is not random, in that the sample of elements that is replaced is fixed for each presentation of a specific compound, but is statistically independent in the sense that the set of elements undergoing replacement is not mutually exclusive for each compound. Thus no assumptions need be made about what the elements code for, nor are there restrictions on the number of elements undergoing replacement.

The replaced elements model accounts for negative patterning and biconditional discriminations in the same way that the

Rescorla-Wagner model accounts for these discriminations – the presence of unique configural elements ensures that the representation of any stimulus compound is distinct from the representations of its constituent stimuli. However, the model is particularly well-equipped to explain other findings concerning differences in responding between single and compound stimuli. In particular, unlike the Rescorla-Wagner model, it provides a comprehensive account of response summation when two CSs are presented in compound. As mentioned above, the evidence concerning summation is mixed. One factor shown to be relevant in determining summation concerns the relationship between the two CSs. Kehoe et al. (1994) observed summation of the conditioned nictitating membrane response in rabbits that had been trained with two CSs from different modalities (one auditory and one visual) but not when the two CSs were both auditory (see also Aydin & Pearce, 1997). This interaction between CS type could extend to include the many failures to observe summation in autoshaping with pigeons (e.g., Aydin & Pearce, 1995, 1997; Rescorla & Coldwell, 1995) since the CSs used in those experiments are from the same (visual) modality. Like the Rescorla-Wagner model, the replaced elements model is readily able to predict summation. However, unlike the Rescorla-Wagner model, it also predicts that the amount of summation observed when two CSs are presented in compound should vary depending on how many elements undergo replacement – the larger the number of elements undergoing replacement, the smaller the amount of summation (Wagner, 2003). Indeed, if 50% of elements are replaced, then no summation should be observed. Therefore, any property that affects the amount of replacement between stimuli should impact on summation. According to Wagner (2003, also Myers, Vogel, Shin, & Wagner, 2001), one such property is whether the stimuli belong to the same or different sensory modalities. Stimuli in the same modality are assumed to inhibit more of each other's elements (i.e., undergo greater replacement when compounded) than stimuli from different modalities. This provides a ready explanation for the failures to observe summation between two auditory CSs in rabbit nictitating-membrane conditioning and between two visual stimuli in autoshaping with pigeons. It also provides an explanation for conflicting data reported by Pearce, Aydin, and Redhead (1997) and Myers, Vogel, Shin, & Wagner (2001). Both groups conducted an experiment that compared responding to a triple compound, ABC, between animals trained with each CS individually (A+ B+ C+) and animals trained with the same CSs as three pairwise compounds (AB+ AC+ BC+). In an autoshaping experiment with pigeons, Pearce et al. (1997) observed summation of responding to ABC in pigeons trained with the two-CS compounds but not in pigeons trained with the three single CSs. In contrast, in an eyelid conditioning experiment with rabbits, Myers et al. (2001) observed greater summation in rabbits trained with the single CSs than rabbits trained with the compounds. Myers et al. suggested that the key factor distinguishing the two experiments was the similarity between the stimuli, since their stimuli were from different modalities (one visual, one auditory, and one vibrotactile) whereas Pearce et al. used all visual stimuli. As Wagner (2003) has shown, the replaced elements model can account for both sets of findings by assuming that the stimuli used by Pearce et al. underwent substantial replacement when compounded (because they were in the same modality) whereas the stimuli used by Myers et al. underwent little replacement.

Further challenges for elemental models.

Retroactive interference in feature negative discriminations. Negative patterning and biconditional discriminations are not the only challenges facing elemental models. Pearce and colleagues have amassed empirical evidence that contradicts the basic elemental approach, including that of the Rescorla-Wagner and replaced elements models that incorporate the notion of a unique configural element in the representational framework. One piece of evidence concerns the sensitivity of feature negative discriminations to retroactive interference. Pearce and Wilson (1991; Wilson & Pearce, 1992) trained animals on an A+ AB− discrimination, and then followed this with B+ training. The B+ training disrupted the previous discrimination performance but did not abolish it, in that the animals continued to respond to A more than to AB. This is not the result predicted by the elemental models described above. According

to the Rescorla-Wagner model, the original feature negative discrimination is solved by acquisition of excitatory associative strength to A and inhibitory strength to B. The subsequent B+ training would imbue B with excitatory associative strength, and thereby reverse performance on the original discrimination (i.e., the animals should respond more to AB than A). The replaced elements model proposed by Wagner and Brandon (2001) also tends to make the same prediction, although this prediction can be reversed if there is a large proportion (i.e. >50%) of elements that undergo replacement between the single stimuli and the compound. However, in this particular case the replaced elements model is constrained to expect much less replacement because the A and B stimuli used by Pearce and Wilson were from different modalities.

The effects of redundant cues on feature negative and negative patterning discriminations. Further evidence against elemental models comes from experiments investigating the impact of irrelevant cues on complex discriminations. In an autoshaping experiment, Pearce and Redhead (1993) compared the performance of two groups of pigeons learning different feature negative discriminations. For one group the discrimination was of the form A+ AB−, for the other group it was of the form AX+ ABX− (i.e., for the second group, X was added to both the reinforced and non-reinforced stimulus configurations). These authors found that the A+ AB− discrimination was learned faster than the AX+ ABX− discrimination. Pearce and Redhead (1993) extended their investigation of the effects of a redundant cue, showing that it also impairs mastery of negative patterning. Rescorla (1972) had previously made a similar demonstration: An AX+ BX+ ABX− discrimination is learned more slowly than an A+ B+ AB− discrimination.

The detrimental effects of the redundant cue on feature negative and negative patterning discriminations are important because, as Pearce (1994) points out, the elemental approach inherent in the Rescorla-Wagner model makes the exactly opposite predictions, as does the replaced elements model of Wagner and Brandon (2001). These models predict that the added cue, X, will facilitate the discriminations. They make this prediction as a consequence of their assumption that the associative strengths of stimuli sum when the stimuli are presented in compound. For example, in the feature negative design, these elemental models anticipate faster acquisition of responding to a compound CS (AX+) than to a single CS (A+) because the former has twice as many elements. As a consequence, they also predicts that B will acquire inhibitory strength faster in the ABX− compound than in the AB− compound because the negative discrepancy between λ (= zero) and ΣV is larger in the former case. For negative patterning the sum of associative strengths will be greater when CSs A and B are presented in compound (AB) than when the compounds AX and BX are presented as ABX. As a consequence, animals should take longer to cease responding to AB− than ABX−, the opposite of what was found.

Pearce's configural model

In a radical departure from elemental models, Pearce rejected the notion that associations form between stimulus elements (Pearce, 1987, 1994, 2002). According to Pearce, any stimulus activates a single configural node that represents the entire pattern of stimulation at that time; this node and this node alone has the capacity to become associated with the US. Thus, presentation of a CS (A) will activate a node representing this event (realistically, this node should also contain information about the background and contextual cues, but for present purposes I will consider it to represent A in isolation), and presentation of a compound (AB) will activate a different node representing that event. This process solves negative patterning and biconditional discriminations because each of the different stimulus configurations activates a different configural node that enters into an excitatory or inhibitory association with the US. However, the process is in danger of being too able to solve these difficult discriminations. For example, it must also account for the fact that these discriminations are more difficult than other discriminations involving the same stimuli: Negative patterning is learned more slowly than positive patterning (A− B− AB+; Bellingham, Gillette-Bellingham, & Kehoe, 1985), and a biconditional discrimination is learned more slowly than a component discrimination (AB+ AC+ BD− CD−) in

which one stimulus (A) reliably predicts the US and another (D) predicts no US (Saavedra, 1975). Anticipating this criticism, Pearce (1987, 1994) proposed that any configural node is fully activated by its complete pattern of stimulus input, but is proportionally activated by part of the input pattern. A product rule provides a simple mechanism to describe the similarity between stimulus compounds. For example, A and AB have a similarity index (s) of ½ because their common component, A, constitutes 100% of one pattern and 50% of the other, and the product of these proportions is ½. AB and AC have a similarity index of ¼ because the common component, A, constitutes 50% of each pattern, and so the product is ¼.

The similarity between stimulus configurations effectively determines the difficulty of any discrimination because learning to discriminate between two input patterns is difficult to the extent that the configural node for each pattern is partially activated by the other pattern. Negative patterning is difficult because the non-reinforced compound AB partially activates two nodes (A and B) that are associated with the US, and thus a substantial amount of responding generalizes to the non-reinforced trials. Positive patterning is easier because the non-reinforced stimuli (A and B) partially activate only one node (AB) that is associated with the US. Similarly, a biconditional discrimination is difficult because each compound (e.g., AC) partially activates the nodes for two other compounds (AB and CD) that have the opposite reinforcement contingency; whereas in the simpler component discrimination, each compound partially activates the same two nodes, but only one of these (CD) has the opposite reinforcement history, while the other (AB) has the same reinforcement history. Perhaps not surprisingly, Pearce's (1994) configural model can also explain how a feature negative discrimination can survive the retroactive interference produced by excitatory conditioning of the inhibitory CS (Pearce & Wilson, 1991; Wilson & Pearce, 1992), and the detrimental impact of a redundant cue on feature negative and negative patterning discriminations (Pearce & Redhead, 1993; Rescorla, 1972). Pearce (1987) also points out that his configural model provides an explanation for overshadowing and external inhibition: both are generalization decrements that occur when the configural node activated during conditioning is only partly activated by the stimulus configuration presented at test.

Problems for Pearce's configural model

The work of Pearce and colleagues has identified several key problems for elemental models like the Rescorla-Wagner model. These problems derive from the assumption that responding to a compound is based on the summed associative strengths of all CSs. However, as already discussed, this very assumption means that those elemental models are able to predict response summation between two separately conditioned stimuli. In contrast to the ease with which Pearce's model deals with complex discriminations, summation effects pose a greater challenge. According to that model, a compound composed of two previously conditioned CSs should activate the configural nodes of those CSs to half strength, and so the generalized associative strength should sum to V ($= \frac{1}{2}V_A + \frac{1}{2}V_B$). That is, the compound should produce the same average level of responding elicited by the individual CSs.

Pearce (1994; 2002) has shown that his model can account for summation if one considers the standard procedure for conditioning two CSs as an AC+ BC+ C− discrimination involving the CSs (A and B) and the conditioning context (C). This can lead to summation because it increases the generalized activation between the reinforced configurations, AC and BC, and the non-reinforced configural unit C. Briefly, the AC and BC units are partially activated on C− trials, causing the C unit to acquire inhibitory strength (to cancel the generalized excitation from AC and BC). As a result, partial activation of the inhibitory C unit on AC+ and BC+ trials leads to superconditioning of the AC and BC units. Therefore, on final test, ABC elicits a large response because it strongly activates the superconditioned AC and BC units (to ⅔ each if A, B, and C have equal salience), but activates the C unit relatively weakly (e.g., to ⅓).

As already noted, summation of responding between two CSs has not been observed in every study to have investigated the

phenomenon (e.g., Aydin & Pearce, 1995, 1997; Pearce et al., 2002; Rescorla & Coldwell, 1995), and the similarity between CSs appears to be a relevant factor in determining when summation occurs (Aydin & Pearce, 1997; Kehoe et al., 1994). In this regard, Pearce's configural model can be seen to provide a successful account of summation effects. According to that model, the loss of summation between similar stimuli occurs because there is less generalization between the two CS configurations and the conditioning context, and thus less superconditioning of the CS configural units (Pearce, 2002). However, Pearce's model is troubled by the finding reported by Myers et al. (2001) of greater summation of eyelid CRs to a triple compound, ABC, in rabbits trained with each CS individually than rabbits trained with three two-CS compounds (AB, AC, and BC). Pearce's configural model predicts the opposite result because the compound ABC should activate the configural representations of each of the three two-CS compounds more strongly than it activates the configural representations of the three single CSs. It should be noted that this specific result was observed by Peace et al. (1997) in their autoshaping experiment with pigeons. However, while the replaced elements model can account for the discrepant findings (Wagner, 2003), Pearce's configural model cannot.

Two further problems for Pearce's configural model concern its explanation of external inhibition and overshadowing. The model explains both as generalization deficits: The addition of the novel stimulus to a CS produces external inhibition because it reduces activation of the CS node; overshadowing occurs when two CSs are conditioned in compound because, when either CS is subsequently tested on its own, it only partially activates the configural node representing the compound and so only evokes a weak CR. But this account is troubled by two findings. The first is that the external inhibition is reduced if the added stimulus is rendered familiar (Brimer, 1970; Pavlov, 1927). Pearce's model is not naturally equipped with a means of explaining this result because familiarity would not be expected to change the similarity between the added stimulus and the CS, and so should not affect the extent to which the CS node is activated by a configuration that includes the added stimulus. Nonetheless the model can explain the above finding with the added assumption that pre-exposure to a stimulus reduces its salience and that this depresses the representation of that stimulus in the configural unit. The second and more problematic finding is that external inhibition produces a smaller deficit than that produced by overshadowing (Brandon, Vogel, & Wagner, 2000), a difference anticipated by the replaced elements model (Wagner & Brandon, 2001). Pearce's model, by contrast, predicts equivalent deficits for overshadowing and external inhibition because they constitute symmetrical changes in stimulus configuration between conditioning and test. To explain the observed difference, Pearce's model would have to be revised to create such an asymmetry (e.g., reducing the activation of configural node AB by input A, and increasing the activation of node A by input AB). However, this revision would greatly alter the amount of generalization predicted to occur between compounds and single CSs, and thereby impact substantially on the model's behavior in many discrimination tasks.

Return to a purely elemental approach

The principal objective of this paper is to show how purely elemental models of stimulus representation can overcome many of the shortcomings of previous elemental descriptions. The commitment to an elemental approach can be justified by the conceptual cost associated with the alternative configural approach. First, it is worth noting that all recent models that incorporate configural representations retain the elemental framework. This is obvious with hybrid models, such as the Rescorla-Wagner model and the replaced elements model of Wagner and Brandon (2001), that explicitly invoke both elemental and configural representations. But even Pearce's theory does not substitute configural for elemental representations. Pearce requires that stimulus representations can be deconstructed into elemental nodes in order to operationalize stimulus similarity. Thus, the nub of Pearce's approach is to retain elemental representations but deny them from directly entering the associative process, instead he adds a layer of (configural)

representation that becomes the locus for the associative mechanism. Therefore, in terms of the representational structures they invoke, purely elemental models are more parsimonious than their configural cousins.

More significantly, configural representations blur the distinction between associations and their arguments (the representations between which associations form). Configural representations, by definition, code for the conjunction of two or more stimulus elements, and as such, they implicitly include associative information (e.g., that stimulus A and stimulus B occurred together). Because this information remains outside the associative mechanism invoked to explain Pavlovian conditioning generally, it must rely on an additional and largely undefined associative mechanism. The distinction between associations and representations becomes blurred when the configural representation constitutes a type of "memory trace" of a stimulus pattern (Pearce, 1987, 1994, 2002), something that is used to recognise that pattern on subsequent encounters or can be partially retrieved by a similar pattern of sensory input. In this case, the associative information implicitly coded by the configural representation fulfills a similar function to that served by traditional associations, since both are essentially records of prior experience that guide future behavior. Moreover, models that include both elemental and configural units (Wagner & Brandon, 2001) are at risk of a combinatorial explosion created by the theoretical possibility of configural nodes that encode the conjunction of all pairwise combinations of elements within a stimulus. To avoid this risk, such models need to specify constraints on the circumstances that give rise to configural representations, such as may arise from the nature of the organism's interaction with its environment.

The challenge for a purely elemental approach to stimulus representation is to explain the variety of findings typically thought to be beyond the scope of elemental mechanisms. As reviewed above, those I take to be crucial tests of any elemental model are:

- How animals master negative patterning and biconditional discriminations.
- How a feature negative discrimination can survive retroactive interference caused by excitatory conditioning of the inhibitory CS.
- The detrimental impact of a redundant cue on feature negative and negative patterning discriminations.

These issues have been instrumental in persuading many theorists of the need to accept configural representations. The difficulty they pose for elemental models relates to assumptions about how stimuli combine. A simple additivity rule, such as that applied in the Rescorla-Wagner model, means that the associative strength of a stimulus compound is a linear sum of the associative strengths of its stimulus components. This assumption renders the model unable to apply the exclusive-OR rule required to solve negative patterning discriminations. The solution adopted by Wagner and Rescorla (1972), that a compound of two CSs activates a unique configural cue, is one way to ensure that stimuli combine in a non-linear fashion because the compound of two CSs is more than the sum of its parts. However, as I describe below, there are other non-linear ways that stimuli may combine that do not depart from a strictly elemental framework and that enjoy considerable success in dealing with other issues that can arise in complex discriminations.

There are two further issues that have featured prominently in the debate about elemental versus configural representations, and are thus issues that I will also consider as important tests of any model. They are:

- The summation of responding when two CSs are presented in compound, and the fact that summation is greater between CSs from different modalities than CSs from the same modality.
- The difference in magnitude of response deficit caused by overshadowing versus external inhibition, and why a novel stimulus should be more effective than a familiar one at producing external inhibition.

In the next section, I describe an elemental model of stimulus representation recently proposed by McLaren and Mackintosh (2000; 2002) and how it deals with each of the above issues. The subsequent sections of the paper will focus on presenting a new elemental model and showing how it too deals with these issues. I will also show how this new

model can explain some recent and otherwise troubling findings by Rescorla (2000; 2001; 2002b) that reveal differences in the amount learned about CSs that are conditioned or extinguished together (in compound) if those CSs have different associative strengths prior to their treatment in the compound. These findings are important because they have been taken as evidence against the assumption, at the core of contemporary learning models including the Rescorla-Wagner model, that stimuli conditioned or extinguished in compound suffer a "common fate". I will show how the new elemental model I propose below, unlike the other models reviewed here, can explain these findings while retaining the common-fate assumption.

The McLaren and Mackintosh model

McLaren, Kaye, and Mackintosh (1989), and more recently McLaren and Mackintosh (2000; 2002), have presented a detailed model of stimulus representation based on Stimulus Sampling Theory. This model contains two crucial features that enable it to explain a wide range of findings about stimulus representation, including mastery of conditional discriminations. First, like other elemental models, it attributes stimulus similarity (and thus generalization) to the extent that stimuli share elements in common, but it assumes that even very different stimuli (such as from different sensory modalities) share a large proportion (50%) of their elements in common[1]. Second, the strength to which an element is activated is not a linear sum of the input strength. Rather, the function relating input to activation strength follows a sigmoid curve characteristic of a cumulative Gaussian distribution. As a result, when two stimuli are presented in compound, the activation strength of the many elements they share in common may be greater or less than the sum of their activation strengths in the individual stimuli. This non-linearity in stimulus compounding provides two mechanisms that enable this elemental model to solve conditional discriminations. First, an element that is weakly activated in each individual stimulus can become strongly activated in the compound. Hence this element may effectively function as an added cue in the same way that the configural element incorporated into the Rescorla-Wagner model is used to solve conditional discriminations (McLaren & Mackintosh, 2002). Second, elements that do not change activation strength between the single and compound stimuli can also assist in solving these discriminations (Rescorla, 1972). For example, a negative patterning discrimination involving two stimuli, A and B, with 50% common *X* elements (with fixed activation weight) is solved when the unique *A* and *B* elements have associative strengths of $-\lambda$ and the common *X* elements have an associative strength of $+2\lambda$ (at this point, the net associative strength for *AX* or *BX* is $2\lambda - \lambda = \lambda$, and for *ABX* is $2\lambda - \lambda - \lambda = 0$). The same approach can also solve biconditional discriminations, although the distribution of associative strengths among common and unique elements is far more complex.

For present purposes, the operation of the second mechanism is more interesting than the first because, as noted above, the first mechanism is functionally equivalent to the added configural element hypothesis adopted by Wagner and Rescorla (Wagner & Rescorla, 1972). Therefore, it is the second mechanism that distinguished the McLaren and Mackintosh model from its forebears in dealing with the issues considered in this paper. For example, unlike the Rescorla-Wagner and replaced-elements models, the McLaren and Mackintosh model can explain how a feature negative discrimination of the form A+ AB− survives excitatory conditioning of the inhibitory CS B. The McLaren and Mackintosh model is also able to explain why a redundant cue impairs mastery of a feature negative discrimination (McLaren & Mackintosh, 2002). However, like the Rescorla-Wagner and replaced elements models, it is unable to explain the disruptive effect of a redundant cue on negative patterning. The model predicts that the added cue will have little impact, i.e., animals will master AC+ BC+ ABC− as quickly as A+ B+ AB−. On the one hand, the model assumes there to be a small proportion of elements distinguishing ABC− from AC+ BC+, making this discrimination difficult, compared with the larger proportion of elements distinguishing AB− from A+ and B+. But this advantage for the simpler discrimination is offset because A and B produce considerable summation when combined in the compound

AB, whereas AC and BC produce only modest summation when combined as ABC. The result of these opposing effects is that the model predicts that the two discriminations will be learned at similar rates.[2]

Like other elemental models, the McLaren and Mackintosh model anticipates summation of responding when two CSs are presented in compound, but the summed associative strength is not predicted to be double that of the individual CSs. This is because the common elements that increase activation strength between the single CSs and the compound are only weakly activated by the single CSs during conditioning and thus acquire little associative strength. In contrast, the common elements that are strongly activated by the single CSs do not change activation strength in the compound and therefore do not support any summation. Conditioned responding is to a large extent controlled by this latter class of common elements because they acquire the majority of associative strength (being present on each conditioning trial, whereas the unique elements are reinforced only half as often). This mechanism explains why summation should be reduced between CSs from the same modality since the proportion of common elements should be greater for such CS combinations (McLaren & Mackintosh, 2002). The model also predicts greater summation of responding to a triple compound, ABC, if the component stimuli have been conditioned individually (A+ B+ C+) than if they have been conditioned in paired compounds (AB+ AC+ BC+), and is thus consistent with the findings reported by Myers et al. (2001) with rabbit eyelid conditioning. Moreover, like the replaced elements model, the McLaren and Mackintosh model predicts this effect will be sensitive to the number of elements shared in common by the CSs, with the difference decreasing as the number of common elements increases. However, the McLaren and Makcintosh model cannot predict a reversal of this effect – greater summation for ABC following AB+ AC+ BC+ training than following A+ B+ C+ training – as has been reported with pigeon autoshaping by Pearce et al. (Pearce et al., 1997).

The McLaren and Mackintosh model anticipates external inhibition for exactly the same reasons as traditional Stimulus Sampling Theory – the novel stimulus adds to the number of stimulus features that can be sampled, and so reduces the probability that the conditioned elements of the CS will be sampled, thereby reducing the CR. Familiarity with the added stimulus could be seen to reduce this effect by reducing the salience of the added stimulus, thereby reinstating the likelihood of sampling CS elements. However, unlike the replaced elements model and Pearce's configural model, the McLaren and Mackintosh model makes no a priori prediction about the relative magnitudes of external inhibition and overshadowing because these are attributed to entirely independent mechanisms (variations in stimulus sampling, and the operation of the delta rule). Therefore, the demonstration that overshadowing produces a greater response deficit than external inhibition (Brandon et al., 2000) is neither support for nor evidence against this model.

A new elemental model emphasizing interactions among stimulus elements.

I now describe a new model of stimulus representation that emphasizes interactions between elements. This model takes as it starting point key features of Stimulus Sampling Theory and the Rescorla-Wagner model. It departs from those models in its description of the processes that govern activation of stimulus elements.

In certain respects, the model is most similar to a modified Stimulus Sampling Theory considered (and dismissed) by Pearce (1994) and Wagner and Brandon (2001). In that approach, the elements of different stimuli inhibit each another's activation in such a way as to hold constant the number of elements activated by any stimulus pattern. For example, if two stimuli, A and B, each activate n elements when presented separately, presentation of the AB compound would activate $\frac{1}{2}n$ of A's elements and $\frac{1}{2}n$ of B's elements. Wagner and Brandon (2001) make the point that this "inhibited elements" approach is in many instances equivalent to Pearce's configural model in terms of the computations that describe similarity between stimulus patterns. However, this inhibited elements model lacks a clear mechanism that would determine which elements of a stimulus

become active in different instances (e.g., which ½*n* of A's elements are activated by the compound AB and which are inhibited). Any adequate description of the process is made even more difficult by the requirement that those elements of stimulus A that are activated by AB must be a statistically independent sample from the A elements activated by the compound AC. The requirement of statistical independence means that the fate of an individual element cannot be linked to any property of the element itself (e.g., its salience), yet it cannot be randomly determined from trial to trial (as could occur in a stochastic sampling process) because the same ½*n* elements of A must be activated every time by the AB compound.

Like the inhibited elements model just described, the model presented below assumes that the elements of different stimuli interact to affect one another's activation. However, rather than preventing that activation, elements of one stimulus effectively reduce the activation of other elements. Importantly, this process is identified with a property of the elements themselves – their activation strength – and is attributed to the operation of a limited-capacity attention buffer, a mechanism that has proved popular in previous elemental models (e.g., Bush & Mosteller, 1951b; Sutherland & Mackintosh, 1971; Wagner, 1981). The core features of the model are laid out below.

1) *A stimulus activates a population of elements, corresponding to different micro-features of the stimulus*. The different elements that constitute a stimulus are activated to different levels (weights) corresponding to the salience of that feature in the stimulus. Similarity between two stimuli is a function of the proportion of elements they activate in common, and these common elements are activated more strongly when the two stimuli are presented together as a compound. For simplicity, I assume that the activation weight of the common elements in the compound equals the sum of their activation weights in the individual stimuli, however non-linear combination rules are equally possible[3].

2) *Activated elements compete for entry to a fixed-capacity analyzer ("attention buffer") as a function of the <u>change</u> in their activation weights*. An element that receives a large increase in its activation can displace a weakly activated element from the attention buffer, or prevent that element entering the buffer. The attention buffer functions as a gain control, increasing and prolonging the activation of the elements it contains. So that this can be operationalized in simulations, I will assume that the buffer doubles the increase in an element's activation weight, although this is obviously a relatively arbitrary magnification factor. For example, if an element's weight increases from 0.1 to 0.4, and this difference is above buffer threshold, the element's weight becomes 0.7. Elements outside the buffer can nonetheless be conditioned and contribute to behavior, but their influence is weaker because their activation weight is not boosted. Figure 2 illustrates the distribution of elements within and outside the attention buffer, and gives an example of how this distribution might change when a stimulus is presented alone versus in compound with another stimulus.

3) *The capacity of the attention buffer is defined by the sum of activation weights, <u>not</u> the number of elements*. Therefore, both the threshold for entry to the buffer and the number of elements in the buffer will vary depending on the average activation weight of its elements. That is, for any array of activated elements ranked in descending order of weight, buffer threshold equals the weight of element *n* (ω_n) for which

$$\sum_{i=1}^{n} \omega_i \leqslant \text{buffer capacity}$$

The change in content of the buffer between single and compound CSs is best illustrated by an example. Consider two equivalent stimuli that each activates 20 elements into the attention buffer when presented alone. In this case the threshold is just below the activation weight of the weakest of the 40 elements. When the two stimuli are presented simultaneously, only a subset of the elements (the strongest ones from each stimulus) will enter the buffer, and the total number will now be smaller than 20 – if the weights are randomly distributed according to a Gaussian density function, there will be on average 16 elements in the buffer, 8 from each stimulus. There are fewer elements in

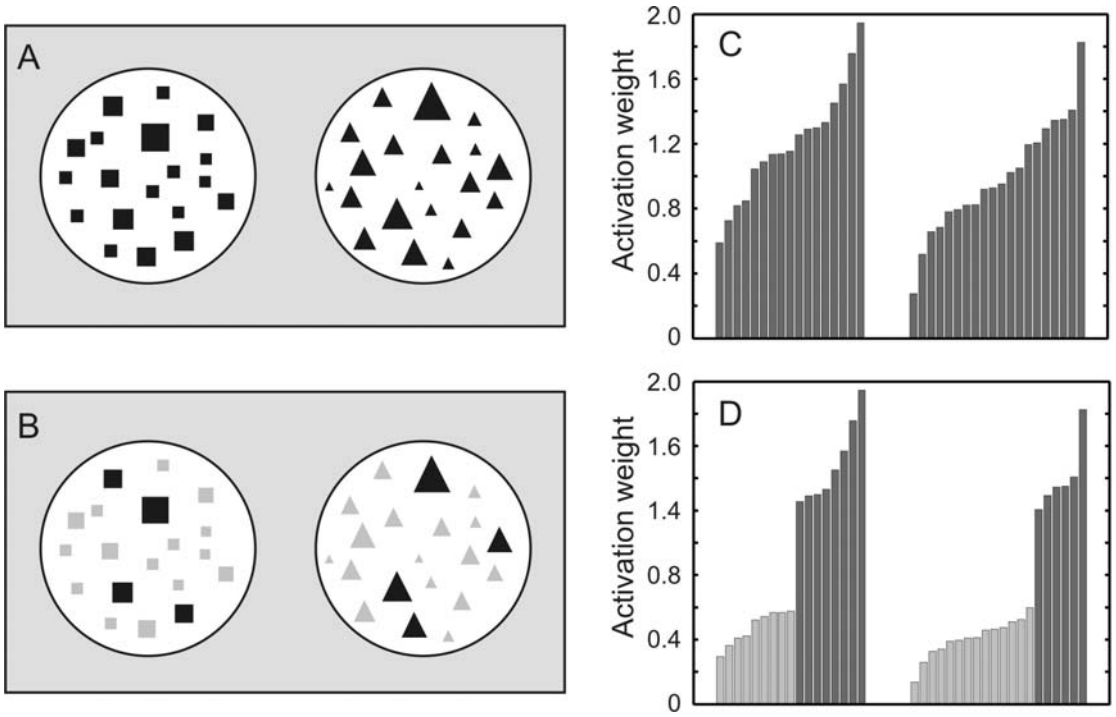

Figure 2. An illustration of the elemental model of stimulus representation proposed here. In **A** & **B**, each small square or triangle represents an element with a particular activation weight (corresponding to its size). Two different stimuli are represented by distinct populations of elements (the squares versus triangles). Depending on the number and weights of other active elements, each element may enter an attention buffer (black shapes) or be displaced from the buffer (gray shapes), according to a rule whereby the summed weight of all elements in the buffer is fixed. **A** shows the two stimuli with all their elements in the buffer (as would be the case if each stimulus were presented on its own); **B** shows the same two stimuli presented together (thus only the more strongly activated elements of each enter the buffer). **C** & **D**: Histograms showing the activation weights for two stimuli, each consisting of 20 elements ordered according to their activation weight (the raw weights are sampled from a Gaussian distribution with a mean of 0.5 and SD of 0.167). **C**: When the stimuli are presented individually, all their elements are activated into an attention buffer which has the effect of doubling their raw weight. **D**: When the two stimuli are presented together, fewer than half of the elements of each stimulus enter the buffer (dark gray bars); most elements (light gray bars) remain outside the buffer without any boost to their raw activation weight.

the buffer because the total weight must remain fixed yet the average activation weight of the elements in the buffer will have increased (because only the stronger elements will have entered). As such, the effective threshold for entry to the buffer will have increased to the value equaling the weight of the 16th element out of the 40. This example serves to illustrate that the threshold does not determine which elements enter, rather the elements that have entered determine the threshold.

4) *Excitatory conditioning occurs when US elements are activated into the attention buffer; inhibitory conditioning or extinction occurs when US elements are activated outside the buffer*. In this regard, the distinction between elements inside versus outside the attention buffer is somewhat analogous to the distinction between the A1 and A2 activation states described in Wagner's (1981) SOP model. However, the current scheme differs from that model in two important regards. First, the role of the buffer in determining excitatory versus inhibitory conditioning applies to the status of US elements only; the status of the CS elements will only affect the rate but not direction of conditioning. Second, the present model does not assume any difference in the rules governing associative changes for excitatory versus inhibitory conditioning. In each case, the change in associative strength is proportional to the increase in activation weight of the US element – a larger increase produces greater excitatory conditioning (if the US element

enters the buffer) or inhibitory conditioning (if the US element remains outside the buffer). As such, changes in associative strength are accurately captured by a delta rule equivalent to that described in the Rescorla-Wagner model (Equation 1). Nonetheless, because US elements have greater activation weight when in the buffer than outside it, the rate of change of associative strength during excitatory conditioning should be greater than during inhibitory conditioning or extinction. Evidence consistent with this has recently been reported by Rescorla (2002a).

5) *During conditioning, the change in associative strength (excitatory or inhibitory) of a CS element is a product of its activation weight (ω).* Thus, ω has the same function as the salience (α) of a CS in the delta-rule, and means that any CS element will be conditioned as long as it is activated (i.e., $\omega > 0$), whether it is inside or outside the buffer. The large effect of the attention buffer on activation weight means that elements inside the buffer undergo much more rapid conditioning than elements outside the buffer. Nonetheless, because elements can undergo some conditioning when outside the buffer, even an associatively activated CS can acquire (or lose) associative strength, as occurs in demonstrations of mediated conditioning or extinction of an absent CS (Hall, 1996; Holland, 1990).

   The above description does not take account of the impact of US elements on activation of CS elements inside the buffer. Given their salience, we can assume that the US elements will displace many CS elements from the buffer, such as might explain the failure to observe effective conditioning in many paradigms with simultaneous CS-US presentations (see Rescorla, 1988 for a review). Nonetheless, the model could be in danger of predicting that conditioning would be very slow with very salient USs, such as shock, because the US would potentially displace all CS elements from the buffer. To circumvent this problem, the model is obliged to allow some temporal slack between the activation of US elements into the buffer and the displacement of CS elements. That is, I assume that displaced elements do not exit from the buffer instantaneously, but follow a decay function, as is often assumed in real-time models of stimulus representation (e.g., Brandon & Wagner, 1998; Kehoe, Horne, Macrae, & Horne, 1993; Wagner & Brandon, 2001). Therefore, while a very salient US like a shock will displace many or even all CS elements from the buffer, the brief overlap of CS and US elements in the buffer should ensure effective conditioning.

6) *An excitatory association between two stimuli means that the elements of one stimulus will activate the elements of the other stimulus; an inhibitory association will suppress that activation.* The associative activation of US elements by a CS is responsible for the CR (e.g., Konorski, 1967). However, as explained in point 7 below, the weight of an element that is associatively-activated is necessarily less than the weight of that element when activated by the stimulus itself. Therefore, although associatively activated elements of a US can elicit a response, this CR will always be weaker than, and potentially of different form to, the unconditioned response elicited when those elements are fully activated by the US itself.

7) *Because excitatory conditioning is proportional to the increase in activation of US elements, and depends on those elements entering the attention buffer (point 4), conditioning is reduced to the extent that the US elements are associatively activated by the CS.* This point reiterates the delta rule, but also explains why the activation of a US element by a CS will always be weaker than the activation of that element by the US itself. This is necessarily the case because the CS-US association will only increase as long as the US elements gain entry to the attention buffer. The US elements will cease entering the buffer, and conditioning will stop, as soon as the CS activates them to a level (ω_L) above which the US itself cannot provoke sufficient increase in their weight to exceed the buffer threshold (i.e., when $\omega_{US} - \omega_L$ = buffer threshold). This means that ω_L represents the maximum excitatory conditioning that can be supported by a US element, and therefore, in the terminology of the delta

rule, $\omega_L = \lambda$. By combining these two statements, we can say

$$\lambda = \omega_{US} - \text{buffer threshold}.$$

Because a CS will activate the elements of a US relatively weakly, those elements are unlikely to enter the attention buffer, and thus would not normally support excitatory conditioning (but would provoke inhibitory conditioning or extinction). However, the description just offered does anticipate an exception to this rule: An extensively trained CS could prime US elements into the attention buffer if the self-generated activation weight of the US element (ω_{US}) is more than twice the buffer threshold (i.e., λ > threshold), as could occur with either a very strong US or a low threshold. This provides a means of explaining the demonstration by Dwyer, Mackintosh, and Boakes (1998) of de novo flavor-preference conditioning between an associatively activated CS (a flavor) and an associatively activated US (sucrose). The reader might recognize that, by permitting a CS to activate US elements into the attention buffer, this undermines the mechanism by which that CS-US association could be extinguished. In this regard, it is pertinent that a characteristic of conditioned preferences for flavors paired with sucrose is their substantial resistance to extinction (Harris, Shand, Carroll, & Westbrook, 2004). In other words, by attributing these effects to a common cause, the model anticipates that behavioral paradigms that can support excitatory conditioning with an absent US will also show resistance to extinction of the primary CS. However, it is not clear at this stage why the flavor-preference paradigm should possess these particular properties.

8) *Just as associations are strengthened between elements of different stimuli (e.g., between CS and US elements), they are also strengthened between elements within a stimulus following the same associative rules*. This provides a further means by which stimulus elements can influence each other's activation. This is important because there are many demonstrations that prior exposure to a stimulus affects learning. Latent inhibition is the cardinal example: conditioning to a CS is retarded if animals have been extensively exposed to that stimulus prior to conditioning (Lubow, 1973). At the same time, familiarity improves discrimination between stimuli (Hall, 1991; Mackintosh & Bennett, 1998). Thus simple exposure to a stimulus leads to changes in the way the stimulus is represented.

In the current model, when a very familiar stimulus is presented, the elements activated by the onset of the stimulus could associatively prime the later elements, thereby preventing those later elements from gaining entry to the attention buffer. This would reduce any response elicited by the stimulus (i.e., cause habituation) and decrease the ability of the stimulus as a whole to undergo conditioning (i.e., produce latent inhibition; see McLaren et al., 1989; McLaren & Mackintosh, 2000 for a detailed description of this process). Conditioning could be reduced even further if the stimulus elements are associatively primed by the context in which the stimulus has been presented previously (Wagner, 1981), giving rise to context-specific latent inhibition (e.g., Lovibond, Preston, & Mackintosh, 1984; McLaren, Bennett, Plaisted, Aitken, & Mackintosh, 1994; Westbrook, Jones, Bailey, & Harris, 2000). Note that this predicts slower conditioning to a pre-exposed CS, but not a failure of conditioning, because conditioning can proceed slowly when CS elements are activated outside the attention buffer. Further, the model enables post-conditioning manipulations (such as testing in a different context) to recover some responding to a pre-exposed CS because it assumes latent inhibition to be a combination of an acquisition deficit and a performance deficit (both caused by the associative priming of CS elements). Therefore, post-conditioning manipulations that oppose the performance deficit can still increase responding (e.g., Westbrook et al., 2000).

With regard to perceptual learning, the model permits familiarity to improve discriminability between similar stimuli via two mechanisms proposed by McLaren and Mackintosh (McLaren et al., 1989; McLaren & Mackintosh, 2000). The first of these mechanisms relies on greater latent

inhibition of their common elements than their distinctive elements. This would arise because the common elements are exposed twice as much as the distinct elements, and therefore greater associative priming would develop among the common elements, and those elements would also receive greater associative priming by the context. The second mechanism is the development of mutual inhibition between the distinctive elements of the similar stimuli. Briefly, once associative links have formed between the common and distinctive elements of each stimulus, presentation of one stimulus will associatively activate the distinctive elements of the other stimulus via their common elements. Inhibitory associations will thus form between the distinctive elements because the distinctive elements of the presented stimulus are active in the buffer while the distinctive elements of the other stimulus are active outside the buffer.

9) *Pre-existing links connect elements, and the strength of these connections changes across the course of conditioning or extinction* (i.e., they would normally have an initial value of zero). An important detail for any connectionist model is to specify constraints on the connectivity between elements. The simplest position is to assume there are no constraints – that all elements are equally inter-connected, and thus any element can be associated with any other element. Although simple, a strong version of this assumption is implausible. Despite the extensive interconnectivity of the nervous system, it is not exhaustive (each neuron is not connected to every other neuron; or, at an even more restricted level, each cortical column is not connected to every other column). Even if each functional unit, corresponding to an individual element, were connected to every other unit, the pervasiveness of variability in biological systems would guarantee that the connections would not be homogenous. Therefore, a more realistic approach assumes variability either in the distribution or the effectiveness of connections between elements. Although this assumption has little bearing on the mechanics of the model as it applies to most situations considered here, it becomes important in accounting for Rescorla's (2000; 2001; 2002b) recent observations that CSs conditioned or extinguished in compound may acquire or lose associative strength at different rates. This important topic is taken up later in the paper.

   As presented here, the model assumes partial connectivity between elements – each element is connected to a subset of the total number of elements (both across and within stimuli). This is comparable to assuming complete interconnectivity between elements but with variability among those connections in terms of their effectiveness to support associations (e.g., variations in β). Differences in connectivity may constitute a means of explaining differences in the rate of conditioning for different CS-US combinations. The textbook example of this is that rats acquire aversions to flavors paired with illness much more readily than lights or noises paired with illness, and conversely learn to fear lights and noises paired with shock much more readily than a flavor paired with shock (Garcia & Koelling, 1966). This apparent bias to learn certain associations may reflect variations in the connectivity of the nervous system – olfactory and gustatory centers may be more extensively connected to visceral centers than nociceptive centers, while auditory and visual centers might be more extensively connected to nociceptive centers than visceral ones.

Based on the assumptions outlined above, I have conducted computer simulations to provide a more quantifiable measure of how the model behaves in circumstances where it is difficult to work through the mechanics of the model using a purely verbal description of the process. To conduct these simulations, further details and assumptions must be made explicit, as described below.

Each single stimulus (CS or US) is represented by 20 elements of varying weight – that is, when the stimulus is presented in isolation, it activates 20 elements into the attention buffer. This is an obviously arbitrary number, but I have confirmed that the simulated results are comparable when using larger or smaller numbers of elements. For convenience, the activation weight of each stimulus element is sampled randomly from a Gaussian

distribution with a mean of 0.5 and SD of 0.167 (thus confined to a range from 0 to 1). This means that CS and US elements were given equivalent weight – not a realistic assumption for many conditioning paradigms, but of little importance here since the model does not attempt to account for interactions between CS elements and US elements. Unless otherwise specified, the different stimuli used in the simulations have no elements in common. Where two stimuli share common elements, the compound of those stimuli activates each common element to a weight equal to the sum of its weights as activated by the individual stimuli. If an element's activation weight increases by an amount that is greater than the buffer threshold, the increase is doubled.

In most instances, I have run multiple simulations with different proportions of interconnectivity between elements (from 10% to 100%). However, there are seldom meaningful differences between simulations generated assuming different levels of connectivity, other than the rate at which conditioning proceeds. Therefore, I will report here only the results of those simulations where the interconnectivity was set at 50% (i.e., each of the 20 CS elements is on average connected to 10 US elements and/or 10 elements of any other CS, including itself). Also, rather than assuming a fixed level of interconnectivity (that each CS element is connected to exactly 10 US elements), I adopt a simpler mechanism that incorporates variability between elements in the extent of their interconnectivity. I have operationalized this by giving each element a fixed probability (0.5) of being connected to any other element.

Equation 2 was used to calculate changes in associative strength (ΔV) between element *x* (e.g., of a CS) and element *y* (of a US) on a trial-by-trial basis. For trials on which feature Y is present,

$$\begin{aligned} &\text{let } \Delta\omega_y = [\omega_y - \sum_{i=1}^{m}(\omega_i \cdot V_{i\text{-}y})] \\ &\quad \text{if } \Delta\omega_y \geq t \\ &\quad \Delta V_{x\text{-}y} = \omega_x \cdot \beta_y \cdot 2 \cdot \Delta\omega_y \qquad (2) \\ &\quad \text{else } \Delta V_{x\text{-}y} = \omega_x \cdot \beta_y \cdot (-\Delta\omega_y) \end{aligned}$$

Note that the associative strength between *x* and *y* ($V_{x\text{-}y}$) increases as long as $\Delta\omega_y$, the difference between the self-generated weight of *y* (ω_y) and the associatively-activated weight of *y*, is greater than the buffer threshold (t), and that ΔV is proportional to twice this difference (as per the effect of the buffer). Otherwise, $V_{x\text{-}y}$ decreases proportional to ω_y (and not twice ω_y because *y* is not in the buffer). Therefore, during normal conditioning, $V_{x\text{-}y}$ will oscillate around its asymptote as $\Delta\omega_y$ alternates between a value above and below t. However, if $\Delta\omega_y$ is well below buffer threshold, as would occur when two extensively-conditioned CSs are presented in compound and reinforced, this could lead to a sustained decrease in $V_{x\text{-}y}$ despite the presence of Y, giving rise to an over-expectation effect (Lattal & Nakajima, 1998).

For trials on which feature Y is absent (i.e., $\omega_y = 0$), $V_{x\text{-}y}$ decreases according to Equation 3

$$\Delta V_{x\text{-}y} = \omega_x \cdot \beta_y \cdot [0 - \sum_{i=1}^{m}(\omega_i \cdot V_{i\text{-}y})] \qquad (3)$$

In all simulations, the learning rate parameter β was set at a nominal value (0.01), selected to give appropriate control over the rate of change in performance.

Finally, all simulations give the expected level of responding to the CS or compound CS on each trial. Thus responding is determined by the sum of the products of the activation weights of the elements and their current associative strength. That is, for stimulus A with *n* elements, responding (R) is given by

$$R(A) = \sum_{i=1}^{n}(\omega A_i \cdot VA_i) \qquad (4)$$

where VA_i is the sum of Vs of each connection from A_i to the US elements. Similarly, responding to a compound, AB, is the sum of $\omega_i \cdot V_i$ for all elements of A and B, but in this case ω is reduced by half for most of the elements of each CS. The plots for all simulations presented in this paper are the average of at least 20 separately run simulations.

To illustrate how associative learning is simulated by the proposed model, Figure 3 shows the acquisition of conditioned responding in a simple Pavlovian conditioning preparation. Equations 2 and 3 were used to calculate the change in strength of the connections between the elements of a CS and US. Conditioning was set in a "context" that comprised twice as many elements as each CS but with the same average activation weight as the higher CS. These context elements competed with the CS and US elements for entry to the attention buffer. Responding, calculated according to equation 4, is seen to increase according to a standard monotonic curve, and the rate of acquisition is faster for the more salient CS.

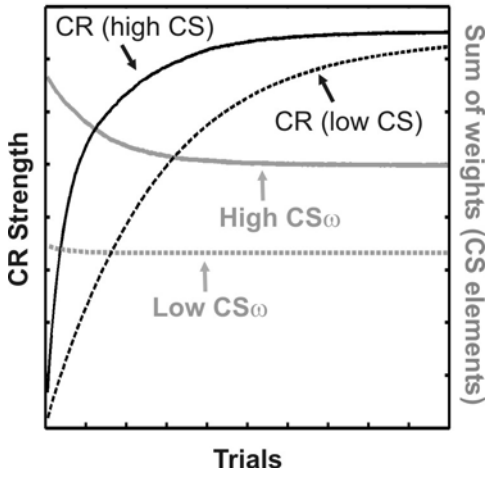

Figure 3. Simulation of a simple Pavlovian conditioning preparation using the elemental model proposed in the present paper. The black lines shows the predicted strength of the conditioned response (CR) to two CSs that differed in salience – the activation weights of the elements of the "high CS" were 50% greater than those of the "low CS" (separate simulations were run for each CS). The figure also shows the sum of activation weights of the CS elements (CSω; gray lines). Across the course of conditioning, the weights decreased as the CS elements became associatively primed by the context.

Figure 3 also shows how the average weight of the CS elements decreased across the course of conditioning as the CS elements became associatively primed by the context (and by other CS elements) thus reducing their entry to the buffer. The decrease in activation obliges the model to predict that responding to the high CS should increase if it were presented in a different context (because, in the new context, the CS elements would not be associatively primed and therefore they would gain greater entry to the attention buffer). However, the fact that this is not typically observed (Hall & Honey, 1990; Harris, Jones, Bailey, & Westbrook, 2000) may be because the predicted increase in responding is offset by the loss of associative strength that had been acquired by the conditioning context itself.

Negative patterning and biconditional discriminations

The mechanisms by which elements interact to influence each other's activation equip the current model with the means to solve negative patterning discriminations (see Figure 4 for a simulation). Many elements of each stimulus (constituting exactly half the total activation weight) are active in the attention buffer when the stimulus is presented alone but are displaced when the two stimuli are presented in compound. Thus the A+ B+ AB− discrimination can be thought of as *Aa*+ *Bb*+ *AB*−, where *a* and *b* represent the weaker elements that only enter the buffer on single stimulus presentations, and *A* and *B* are the stronger elements active in the buffer during both single and compound stimulus presentations. It should be clear that *a* and *b* will ultimately acquire all of the associative strength. Ordinarily, there would still be substantial responding on AB− trials because the *a* and *b* elements are active, even though their activation is not boosted by the attention buffer. However, across the course of AB− trials, inhibitory associations would be acquired between the stronger elements of each CS and the weaker elements of the other CS (i.e., from *A* to *b*, and *B* to *a*) because the stronger elements are active in the attention buffer while the weaker ones are active outside it. Thus, mastery of the discrimination depends on two things: (1) the excitatory associative strength becoming confined to the weaker elements; and (2) the inhibition of the weaker elements by the stronger elements of the other CS on the AB− trials. Within the current model, there is also opportunity for unique interactions between stimuli such as might support biconditional discriminations (AB+ CD+ AC− BD−; simulations of this are presented in Figure 5). The opportunity arises because different stimuli are assumed to have different populations of elements and thus

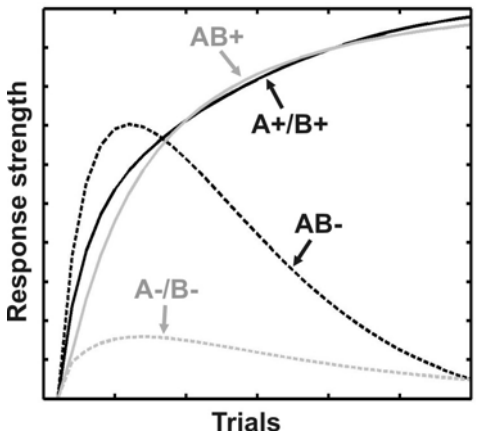

Figure 4. Black lines: Response strength to reinforced stimuli (A+/B+) and a non-reinforced compound (AB-) in a negative patterning discrimination, as simulated by the elemental model proposed in this paper. For comparison, simulated response strength is also shown (gray lines) for single and compound stimuli in a positive patterning discrimination (A- B- AB+).

different distributions of activation weights. Therefore, the threshold for entry into the attention buffer will vary for different compounds. For example, if there are fewer salient elements in stimulus B than C, then there will be some elements of A that are above threshold in the AB compound but are below threshold in the AC compound. While appropriate counterbalancing of stimuli across an experiment will prevent systematic inequalities (i.e., counterbalancing between B and C will ensure that B has fewer salient elements than C for half the subjects, but more salient elements than C for the other subjects), there will always be an inequality in each individual case to support mastery of the discrimination.

The possibility that different stimuli might share elements in common provides another means by which elements could combine in distinct ways in different stimulus compounds. Specifically, if common elements have increased activation weight when stimuli are presented in compound because of the combined input they receive from both stimuli, they would have increased likelihood of entering the attention buffer when stimuli are presented in compound. This would enhance the distinctiveness of different stimulus compounds because the different stimulus combinations should share different elements in common. The right plot in Figure 5 shows how as little as 10% overlap between stimulus pairs can assist the model in solving the biconditional discrimination. This mechanism would also assist with solving negative patterning discriminations by facilitating the acquisition of inhibitory associative strength to the common elements (because these elements would be more strongly activated in the non-reinforced compound than in the reinforced single CSs).

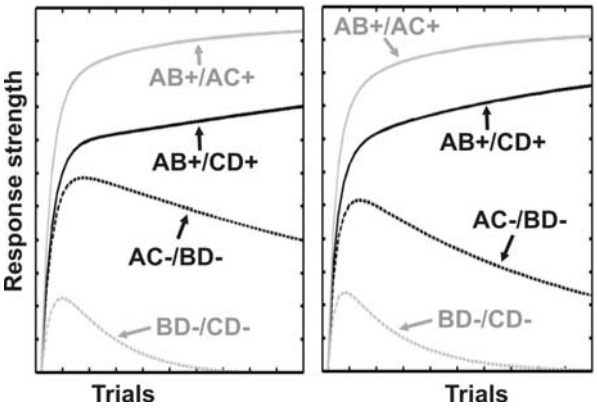

Figure 5. Black lines: Response strength to compound stimuli in a biconditional discrimination (AB+ CD+ AC− BD−), as simulated by the elemental model proposed in this paper. For comparison, simulated responding strength is shown (gray lines) for a component discrimination (AB+ AC+ BD− CD−). In these simulations the four stimuli either had no common elements (left) or shared 10% of their elements in common (right).

Retroactive interference in feature negative discriminations

The current model correctly predicts the finding by Pearce and Wilson that feature negative discriminations can survive retroactive interference when the inhibitory CS undergoes excitatory conditioning (Pearce & Wilson, 1991; Wilson & Pearce, 1992). The prediction emerges from the manner in which excitatory associative strength is distributed among the elements of A and B (again I'll use *A* and *B* to denote the stronger elements of each CS that enter the attention buffer in the AB compound, and *a* and *b* to denote the weaker elements active in the buffer during single stimulus presentations but displaced on trials with the AB compound). Crucially, by completion of the initial feature negative conditioning, most of A's associative strength will have been acquired by the *a* elements, while the *B* elements will have acquired

inhibitory strength against both the US and the *a* elements. When B undergoes excitatory conditioning in phase 2, the majority of its excitatory strength will be acquired by the *b* elements because they start phase 2 with zero associative strength, whereas the *B* elements start with inhibitory strength. In other words, much of B's excitatory associative strength will be reduced when it is again presented in compound with A because the *b* elements will not enter the attention buffer. Further, even though the *B* elements will have lost their inhibitory strength against the US, they will have retained their inhibitory strength against the *a* elements, thus preserving the ability of B to reduce responding to A.

The effects of redundant cues on feature negative and negative patterning discriminations

As shown in Figure 6, the model proposed here correctly predicts that a redundant cue will impede learning of a feature negative discrimination (Pearce & Redhead, 1993). The prediction can be understood by considering the proportion of stimulus weights that are active in the attention buffer during reinforced and non-reinforced trials. With the standard A+ AB− discrimination, only half of A's activation weight is in the buffer during both A+ and AB− trials, leaving the other half to acquire the excitatory associative strength. With the AX+ ABX− discrimination, ⅔ of AX's activation weight is in the buffer during AX+ and ABX− trials, leaving only ⅓ to acquire the excitatory associative strength.

The disruptive effect of an added cue on negative patterning (Pearce & Redhead, 1993; Rescorla, 1972) constitutes a crucial challenge for elemental models since each of the other elemental models reviewed here failed to predict that the AX+ BX+ ABX− discrimination is harder than the A+ B+ AB− discrimination. Unlike those other models, the current model does correctly predict this result, as confirmed by the simulation shown in Figure 7. As described earlier, mastery of an A+ B+ AB− discrimination depends on two things: (1) the excitatory associative strength becoming confined to weaker *a* and *b* elements of each CS, and (2) the inhibition of those *a* and *b* elements by the *B* and *A* elements. But this second process is undermined in the AX+ BX+ ABX− discrimination, and particularly so for the *x* elements (here, because no stimulus is presented alone, I use lower case italicized letters, such as *x*, to stand for elements that are in the buffer during the reinforced compounds AX and BX, but are displaced from the buffer in the non-reinforced triple compound ABX, and I used upper case italicized letters to stand for those elements active in the buffer for both double and triple compounds). The *x* elements do not become inhibited by the *A* and *B* elements on ABX− trials because those same *A* and *B* elements are positively paired with the *x* elements during each AX+ and BX+ trial (i.e., the *A* and *x* elements are co-active in the attention buffer on AX+ trials, and the *B* and *x* elements are co-active in the buffer on BX+ trials). As a result, the *x* elements will continue to support generalized responding on ABX− trials. This generalized responding will be substantial because the *x* elements will take up much of the excitatory associative strength on AX+ and BX+ trials, being reinforced twice as often as the *a* and *b* elements.

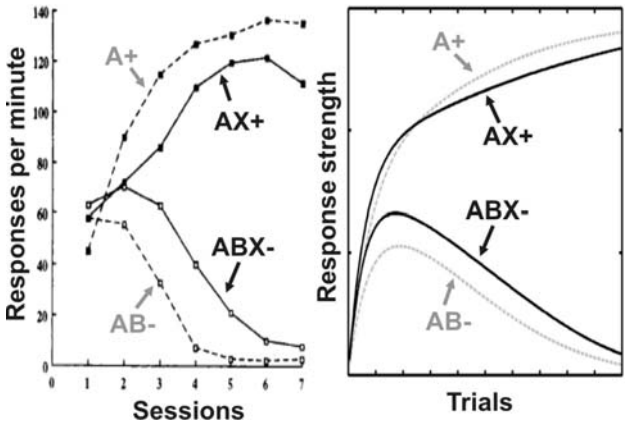

Figure 6. Left (reproduced from Pearce, 1994, with permission of the author): Results of an experiment by Pearce & Redhead (1993) comparing the rates of responding between two groups of pigeons trained on different feature negative discriminations: one group (broken lines) was trained with a standard A+ AB− discrimination, the other group (solid lines) was trained on an equivalent discrimination that included a redundant cue (i.e., AX+ ABX−). Right: Response strength on the same two discriminations, simulated using the model proposed here. Like the pigeons, the model takes longer to discriminate between AX+ and ABX- than between A+ and AB-.

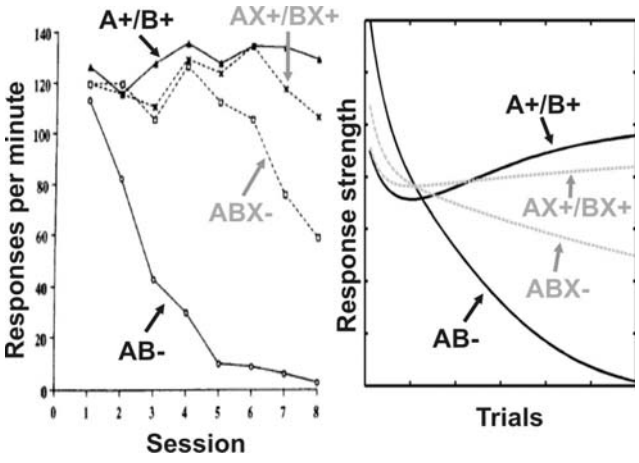

Figure 7. Left: Results of an experiment by Pearce & Redhead (1993, reproduced with permission of the authors) that trained one group of pigeons on a standard negative patterning discrimination (A+ B+ AB−) and a second group on an equivalent discrimination with a redundant cue (AX+ BX+ ABX−). Pigeons in both groups had been pretrained with the reinforced stimuli. Right: Simulated performance on the same discriminations according to the elemental model proposed here. The model captures the basic observation from the experimental study that the standard negative patterning discrimination (black lines) is mastered more quickly than the discrimination with the redundant cue (gray lines).

Summation

The summation of responding when two CSs are presented in compound does not present any difficulty for elemental models, including the present one. The model anticipates summation because all elements from each CS are activated by the compound, even though half the elements are activated outside the attention buffer and will thus make a smaller contribution to responding. The predicted amount of summation is best illustrated by an example involving two CSs, A and B, for which the associative strength is split equally among *A* and *a* elements, and *B* and *b* elements. If the activation weight of the *a* and *b* elements inside the attention buffer is twice their weight outside the buffer, the CR to the compound AB will be 50% greater than the CR elicited by A or B alone (this answer comes about because the CR is the product of V, which for the compound is twice that of the single CSs, and the average activation weight, which for the compound is ¾ that for the single CSs).

As mentioned earlier, there are numerous reported failures to observe summation in Pavlovian conditioning paradigms. Thus any model that is committed to predicting summation has little more empirical support than a model incapable of prediction summation. Clearly the most desirable feature of any model that purports to deal with this topic is to explain both the occurrence and absence of summation and identify the crucial factors that determine which outcome will be observed in a given case. A factor that has already been identified concerns the similarity between the two CSs: Summation is frequently observed between CSs from different modalities, but is not observed between CSs from the same modality (Aydin & Pearce, 1997; Kehoe et al., 1994). The current model can explain this pattern by appealing, once again, to the role played by common elements. Conditioning of two similar stimuli, A and B, would involve alternating *AX*+ and *BX*+ trials (where *X* are the common elements). In addition to excitatory conditioning of *A*, *B*, and *X* elements, this will lead to the acquisition of inhibitory associations between the unique *A* and *B* elements (via the mechanism described earlier in the discussion of perceptual learning, see point 8). As a result each will suppress responding to the other when presented together in compound, thus undermining the basis for response summation. The same process could be engaged by the conditioning context, with the context serving the same role as the common elements in fostering the development of inhibition between the two CSs. This latter mechanism predicts that summation, measured as the difference in response strength to the compound versus the individual CSs, will be greater when two CS are conditioned in different contexts than when they are conditioned in the same context.

The predicted effect of stimulus similarity on summation is illustrated in Figure 8: the left plot shows the amount of summation produced by combining two CSs that have no common elements, the right plot shows the summation produced by two CSs sharing ⅓ of their elements in common. It is clear that the presence of common elements has served to reduce summation. It is also clear that, despite the presence of these common elements, some summation is observed initially, but disappears with extended conditioning. This prediction is

problematic because the available experimental evidence shows that CSs from the same modality produce no summation across the entire course of conditioning (Kehoe et al., 1994). The model makes this prediction when it assumes that the excitatory CS-US associations that support summation are acquired earlier than the inhibitory associations between CS elements that suppress summation. In general, this is a reasonable assumption because the latter inhibitory associations can only be acquired after excitatory associations are acquired between common and distinctive CS elements.

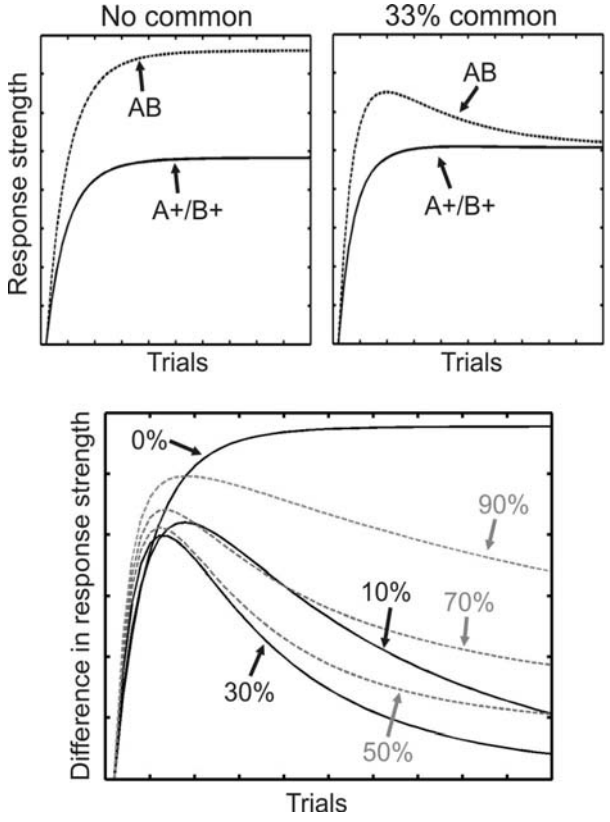

Figure 8. Top: Simulated response strength, generated by the elemental model proposed here, when two separately conditioned stimuli (A+/B+) are presented in compound (AB). The two upper plots show predicted response strength when the two CSs are assumed to share either no common elements (left) or ⅓ of their elements in common (right). Common elements serve to reduce summation by allowing the distinctive elements of each CS to acquire mutually inhibitory links. However, this effect is sensitive to the proportion of common elements, as shown in the bottom plot: Summation (the difference in response strength between the compound and the single CSs) first decreases as the overlap between the CSs increases, but the expected summation increases again as the overlap exceeds 30%.

However, the prediction changes if different parameters are selected for the acquisition of associations among CS elements versus associations between CS and US elements, differences that might arise because of the simultaneous versus sequential nature of the events themselves. For example, for Pavlovian paradigms in which conditioning proceeds slowly, within-CS associations may be acquired early in training and thus be available to suppress summation from the first appearance of the CR. By extension, paradigms that support rapid conditioning should be particularly likely to show summation in early stages of training. Nonetheless, intermixed presentations of the CSs prior to such conditioning would serve to reduce summation by establishing in advance inhibitory associations between the distinctive elements of the CSs.

Figure 8 also shows how the prediction of summation is sensitive to the proportion of overlap between the CSs – summation is suppressed when the two CSs share 50% or fewer of their elements in common, but summation progressively returns if the CSs share a large proportion of elements in common (e.g., 90%). Summation reappears under these circumstances because the mechanism that suppresses summation – mutual inhibition between distinctive elements – affects only a small proportion of the associative strength of the two CSs.[4]

Differences in the development of mutual inhibition between distinctive elements are also relevant to the conflicting results regarding the summation of responding to a triple compound, ABC, when the three stimuli are conditioned individually (A+, B+, and C+) versus as three pairwise compounds (AB+, AC+, and BC+). Pearce et al. (1997) observed summation of keypeck responses to the compound ABC in pigeons trained with the two-CS compounds but not in pigeons trained with the single CSs; whereas Myers et al. (2001) observed greater summation of conditioned eyelid responses to ABC in rabbits conditioned with the three CSs separately than in rabbits conditioned with the two-CS compounds. In general, the current model predicts greater summation when the three CSs are conditioned individually than as paired compounds (consistent with the results reported by Myers et al., 2001). However, this

difference is sensitive to the presence of common elements because, as in the case of simple summation between two CSs, the common elements will serve to establish mutually inhibitory associations between the distinctive elements of the CSs when they are conditioned individually, thereby reducing the amount of responding elicited by the compound ABC. Note that this process will not occur when the CSs are conditioned as paired compounds because any inhibitory associations between the distinctive elements will be prevented every time those elements occur together in one of the compounds. However, as in the case of simple summation, the current models predicts that the above difference will only emerge later in the course of training. In other words, early in training animals trained with the single CSs should show always greater summation than animals trained with compound pairs, whereas with extended training, this difference should disappear and may even reverse if the stimuli share many elements in common.

External inhibition and overshadowing.

Like the replaced elements model and Pearce's configural model, the elemental model proposed here attributes external inhibition to a loss of CS activation (CS elements that would otherwise receive a boost in weight by entering the attention buffer are displaced from it by elements of the added stimulus). The current model is also readily equipped to explain why a familiar stimulus is less able than a novel stimulus to induce external inhibition (Brimer, 1970; Pavlov, 1927). Across the course of exposure to the stimulus, associations would form between its elements (and between the context and the stimulus elements) with the consequence that elements of the stimulus would be associatively primed and thus prevented from subsequently entering the attention buffer. Thus the familiar stimulus places less demand on the attention buffer and so is less effective at producing external inhibition because it displaces fewer elements of the CS from the buffer.

The current model can be seen to predict that external inhibition will decrease as the similarity between the CS and added stimulus increases. For example, the response deficit should be smaller when the two stimuli are in the same modality than from different modalities. This is predicted because an added stimulus from the same modality will have fewer distinctive elements that can displace the conditioned elements of the CS from the buffer. In addition, their common elements will have increased representation in the buffer due to the summed inputs from the two stimuli. Since the common elements will have been conditioned, their increased activation in the compound will serve to maintain the CR. Essentially the same prediction is made by Pearce's configural model and the McLaren and Mackintosh model. In the former case, the more similar the added stimulus and CS, the more their compound will activate the configural node of the CS. According to the McLaren and Mackintosh model, the response deficit is determined by the increased probability that CS elements will not be sampled when the new stimulus is added to the CS. This probability is a positive function of the number of distinctive elements in the added stimulus, and is therefore reduced as the number of common elements increases. In this respect, these three models can be distinguished from the replaced elements model of Wagner and Brandon (2001) which makes the opposite prediction. According to that model, there will be greater external inhibition between stimuli in the same modality because there will be a greater replacement of CS elements when those stimuli are compounded compared to stimuli from different modalities.

The current model can also explain the asymmetry between external inhibition and overshadowing reported by Brandon et al. (2000). When a novel stimulus is presented with a CS, those CS elements that are displaced from the attention buffer remain active (but with smaller weight), and so continue to contribute to responding. In other words, even if half the elements of the CS were displaced by the novel stimulus, the CR would be reduced by only ¼. But the equivalent calculation does not apply in the case of overshadowing: Because the acquisition of associative strength for all elements is governed by a single error term, the elements of each CSs will gain, on average, only half the available associative strength.

The current model also predicts a difference between overshadowing and external inhibition in terms of the effects of stimulus similarity. As described above, there should be less external inhibition when the CS and added stimulus are similar (e.g., from the same modality). By contrast, the amount of overshadowing between two CSs should not be affected by their similarity. When two CSs, A and B, are conditioned in compound, associative strength will be distributed among the distinctive elements of each CS and their common elements. Therefore, the generalization of responding to A or B alone will be reduced by two factors: (1) the amount of associative strength that has been taken up by the distinctive elements of the other CS; and (2) the drop in activation weight of the common elements. While the relative influence of these two factors will change as the similarity increases (the importance of the first factor will decrease and the 2nd will increase), they are effectively complementary such that the net effect on responding remains constant (a conclusion confirmed by conducting simulations of the process). In this regard, the current model can be distinguished from the other models considered here. The replaced elements model predicts that overshadowing increases as the similarity of the CSs increases, because of the greater number of elements that undergo replacement between single and compound presentations. Pearce's configural model makes the opposite prediction – as two CSs become more similar, their ability to activate the compound configural node increases, thus reducing the generalization deficit. Like Pearce's configural model, the McLaren and Mackintosh model predicts less overshadowing between similar CSs because the deficit in responding to either single CS is only reduced to the extent that the distinctive elements of the other CS have acquired associative strength during compound conditioning.

Changes in associative strength when CSs are conditioned in compound.

Much of the discussion thus far has been concerned with how the elements of different stimuli compete for access to an attention buffer and the consequences this has for the CR. I have said comparatively little about the consequences of this competition for conditioning. In the simplest case, when two neutral CSs are presented in compound and reinforced, the outcome is straightforward – their elements share the available associative strength, and those elements that enter the attention buffer acquire much more associative strength than those excluded from the buffer. Thus, if two CSs have equal salience they will acquire equal associative strength, and this will be half the strength acquired by either stimulus if it were conditioned alone. This is a direct effect of the delta rule proposed by Rescorla and Wagner (1972) and adopted by almost all contemporary models of associative learning. But this view has been seriously challenged by Rescorla's recent discovery that CSs conditioned or extinguished in compound do not undergo equal changes in associative strength if their initial strengths differ (Rescorla, 2000, 2001, 2002bb). Below, I describe these findings and show how the current model can explain them.

Kamin's (1968) blocking paradigm is the clearest example of a design in which stimuli with differing associative strengths are conditioned in compound: one stimulus (A) is conditioned before it is combined with a second stimulus (B) and the compound reinforced (i.e., A+ followed by AB+). The significant finding Kamin reported was that B underwent less conditioning during AB+ trials than if A had not been pre-conditioned. The Rescorla-Wagner model readily explained this finding by asserting that changes in the associative strength of any CS is determined by the summed associative strength of all CSs present on the trial. Thus conditioning to A and B during AB+ trials in phase 2 is curtailed by the associative strength already acquired by A in phase 1. Note that, according to the Rescorla-Wagner model, conditioning to A and B is curtailed – if A and B have equal salience, then each will acquire half of what limited associative strength remains available. It is this prediction that is falsified by Rescorla's recent investigations.

Rescorla (2001) showed that, in a blocking paradigm similar to that described above, B acquired greater excitatory conditioning than did A during AB+ trials. (Note, this difference refers only to the change in associative strength that occurred during compound conditioning in phase 2. The terminal associative strength of A would still be greater

than that of B as a result of the prior conditioning of A in phase 1.) The effect was not confined to the standard blocking design in which B is neutral prior to AB+ training; Rescorla (2000) observed the same effect if B had been pre-trained as a conditioned inhibitor (i.e., animals were trained on an A+ X+ BX− discrimination in phase 1, and AB+ in phase 2). Nor did it depend on the same US being used during phases 1 and 2: Greater excitatory conditioning to B than A was also observed if animals had been trained on an A++ AB− discrimination prior to AB+ conditioning (Rescorla, 2002bb). Clearly the inequality in conditioning to A and B depends on their difference in associative strength at the beginning of compound conditioning. Rescorla reported comparable effects when the AB compound was not followed by a US in phase 2 (Rescorla, 2000, 2001, 2002bb). In this case, however, the change (decrease) in associative strength was greater for A than B.

The above findings are particularly difficult for configural theories, like that of Pearce, in which conditioning involves a single associative change because this excludes the possibility of differential changes to the components of a compound. Elemental models, on the other hand, are better equipped to deal with such findings because compounds contain multiple sources of associative change, thus providing the opportunity for differential conditioning of stimulus elements. Nonetheless, conventional elemental accounts, including the Rescorla-Wagner model, are also seriously challenged because these models assume that the same associative change is applied to each element (ie, they assume a "shared fate" for stimuli conditioned or extinguished in compound). As such, the findings have been taken as support for models that use separate error terms to determine the change in associative strength of different CSs (e.g., Mackintosh, 1975), and have encouraged revisions to the Rescorla-Wagner model such as Rescorla's (2000; 2001) suggestion that the change in associative strength of a CS, A, during compound conditioning is determined by the common error term ($\lambda - \Sigma V$) multiplied by that CS's own error term ($\lambda - V_A$).

It is of some significance that the elemental model proposed here can readily explain these recent findings while retaining the principle that changes in associative strength for all stimuli are determined by a common error term. The process by which this occurs is shown in Figure 9, and depends in large part on the assumption (#9) of variability in the connectivity between elements. This assumption is implemented here as random variability in the distribution of connections – two elements have a certain probability of being connected. However, the operations described below would equally apply if one assumed complete interconnectivity between elements but random variability in the efficacy of those connections to support associative learning.

As illustrated in Figure 9, the associative strength of a CS is represented by an array of associations between CS and US elements. Each US element supports a particular level of associative strength, and the strength of any single connection between a CS element and US element will be determined by the activation weight of its CS element relative to the weights of the other CS elements that converge on the US element. When stimulus A is conditioned, its associative strength is distributed equally among the stronger *A* elements (those that enter the attention buffer when A is presented in compound with B) and the weaker *a* elements (those displaced from the buffer in the AB compound). Therefore, when A is presented in compound with B, half of A's acquired associative strength is reduced (to the extent that activation of the *a* elements is now weaker). This means that many US elements can support further conditioning during this second phase, even if conditioning of A had proceeded to asymptote in phase 1. However, the recovered associability is not uniformly distributed across the US elements, but is largely confined to the US elements that receive few connections from the stronger *A* elements. This is because ΣV remains high for the US elements that receive many connections from the *A* elements, so there is less opportunity for further conditioning with those elements. In other words, the distribution of the residual US associability is negatively correlated with the distribution of connections from *A*. In contrast, there is no correlation between the US associability and the connections from *B*. This difference means that B will acquire the greater share of the available associative strength. The same logic explains why the pre-trained stimulus A

should undergo greater loss of associative strength than the novel stimulus B if the compound AB is not reinforced. In this case, the *A* elements are better connected than the *B* elements to those US elements with the strongest associative input (and hence support the greatest decrease in associative strength). These arguments have been confirmed by computer simulations shown in Figure 10.

The account offered above can be extended to generate a novel and testable prediction. As described above, when a CS, A, and a neutral stimulus, B, are presented in compound and not reinforced, the associative strength of A decreases more than that of B because the US activation is preferentially distributed among those US elements with strong input from the *A* elements but is not correlated with the input from the *B* elements. It follows that the opposite result will be observed if the US activation becomes preferentially distributed among those US elements with more connections from the *B* elements than from the *A* elements. In such a case, there should be a greater decrease in associative strength for B than A. One way to achieve this uses an over-expectation design (Lattal & Nakajima, 1998). Animals are first trained with a compound, AB, followed by further conditioning of one of the CSs, A, before returning to conditioning of the compound (i.e., AB+ then A+ then AB+). The return to AB+ conditioning in phase 3 should give rise to an over-expectation effect because the elements of A (*A* and *a*) should have reached asymptote (λ) by the end of phase 2 while the *B* elements should also have associative strength of ½λ from conditioning in phase 1. Crucially, however, the "excess" associative strength acquired during A+ training in phase 2 will be preferentially distributed among the US elements that receive more connections from the elements of B than A. This will necessarily occur because the associative strength that becomes available

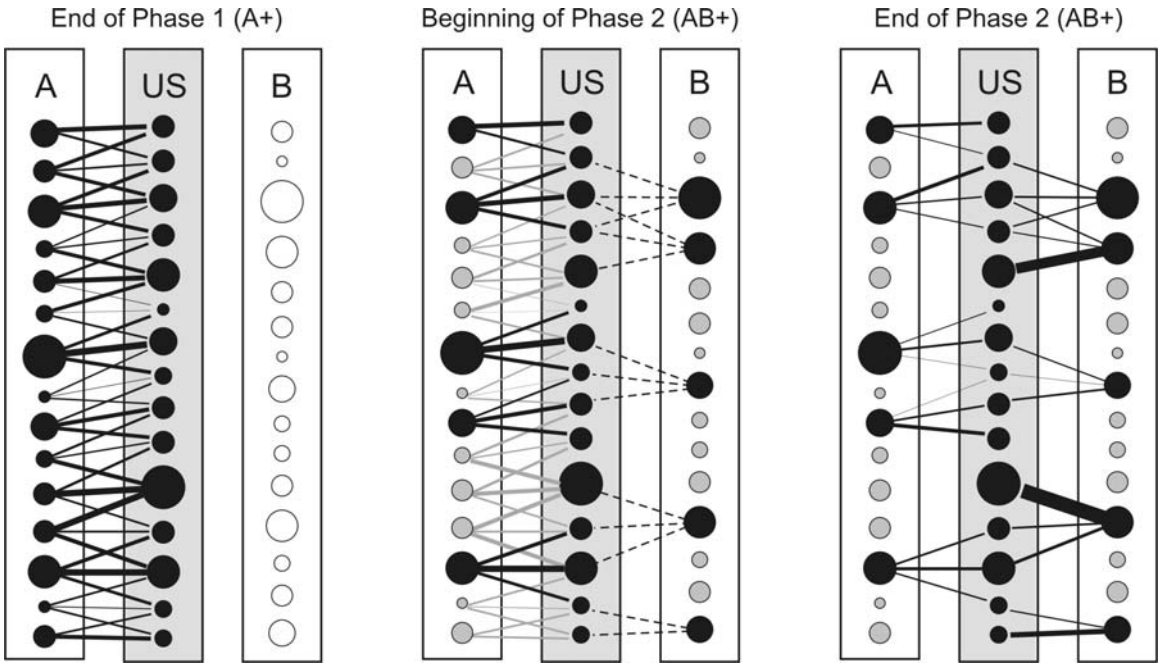

Figure 9. Illustration of how the associative strength between two CSs (A and B) and a US changes across phases of a blocking paradigm, according to the elemental model proposed here. The stimuli are comprised of 15 elements (circles) of varying activation weight (corresponding to their size). Each CS element is connected to 3 US elements. Across conditioning, the strength of each connection grows as per the error-correction rule proposed by Rescorla and Wagner (1972). The left diagram depicts the connections between A and the US after A+ conditioning in phase 1. The center diagram shows the associative connections at the beginning of phase 2 when A and B are presented together. Half of the previously established associative strength is reduced because many A elements are displaced from the attention buffer (shown as gray circles) effectively halving their activation weight. As a result, many US elements are able to support further conditioning in phase 2. But because these US elements receive few connections from the A elements that remain in the attention buffer, the associative strength acquired de novo during phase 2 is biased towards elements in B, as shown in the diagram on the right.

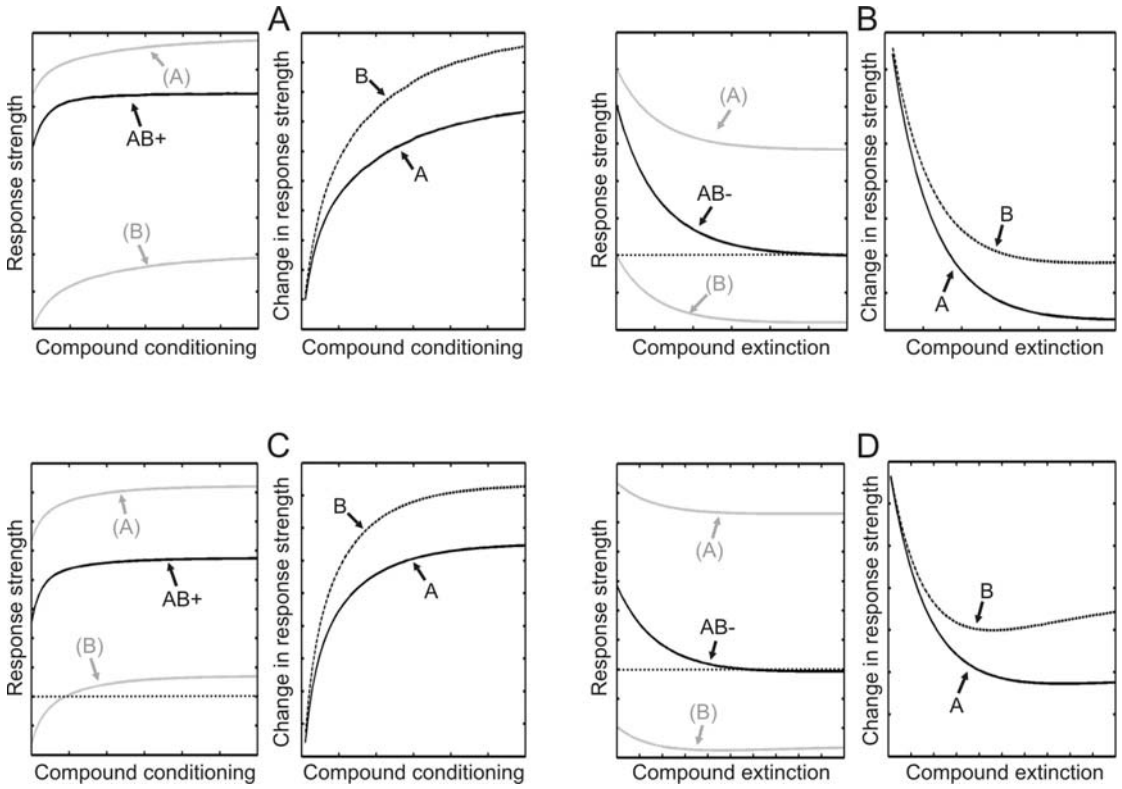

Figure 10. Response strength, as simulated by the elemental model proposed here, for four experimental designs investigated by Rescorla (2000; 2001; 2002b). **A**: The left plot shows the predicted response strength during conditioning of a compound, AB, following pre-conditioning of CS A. The gray lines show the response strength that would be elicited by CS A or CS B alone. The right plot shows the predicted change in response strength to those CSs during the compound conditioning (i.e., their response strength across compound conditioning minus their response strength at the beginning of that conditioning). **B** shows predicted responses for the same design as A except that responding to the compound is extinguished (AB-). **C** and **D**: The left plots show predicted response strength to CS A and CS B during conditioning (C) or extinction (D) of the compound, AB, following pre-training on an A+ X+ BX- discrimination. The right plots show the change in response strength to CS A and CS B. The dotted horizontal lines in B, C, and D mark zero response strength – negative values indicate inhibitory conditioning.

for conditioning in phase 2 is distributed in large part among those US elements that had received strong input from the *B* elements in phase 1. Therefore, these same US elements will be "over-activated" by the AB compound in phase 3, and as a result will produce the greatest decrease in associative strength. Thus B will lose associative strength faster than A in phase 3, a prediction I have confirmed with computer simulations.

Concluding remarks.

The present paper has reviewed previous elemental and configural models of stimulus representation and presents a new elemental model based on the approach laid out by Stimulus Sampling Theory. The new model differs from previous elemental models in its assumptions about how stimulus elements interact for entry to a limited capacity analyzer (attention buffer) and its description of the fate of elements excluded from the attention buffer. I have shown how this new model can explain a large number of behavioral findings that previously have been taken as evidence contradicting the basis of the elemental approach. Most importantly, it permits a solution to negative patterning and biconditional discriminations without appealing to notions of configural representations. The current model also fares better than most elemental models in explaining a variety of findings by Pearce and colleagues concerning the detrimental impact of a redundant cue on negative

patterning and feature negative discriminations. This is important because these findings have been used to argue for the need to invoke the notion of configural representations in addition to elemental representations. Another advantage of elemental models, such as the one presented here, is that they are well-equipped to explain how stimulus representations change with experience, thus providing a mechanism by which familiarity with a stimulus may reduce its associability (i.e., latent inhibition).

Finally, by allowing elemental stimulus units to enter directly into the associative process, the current model can explain recent findings about differences in the fate of stimuli conditioned (or extinguished) in compound. These findings have been interpreted as evidence against the assumption, at the core of the Rescorla-Wagner model, that stimuli conditioned in compound undergo the same change in associative strength. The explanation provided by the current model retains that core assumption, and attributes the evidence for differences in the rate of conditioning to specific biases in the distribution of available associative strength across US elements.

Ultimately, the worth of any proposed scheme for stimulus representation will depend not only on its explanatory power, but how satisfactorily the mechanisms it invokes can be described and tested. The model proposed here explicitly avoids invoking configural representations. Instead, it uses a limited-capacity attention buffer as the mechanism responsible for the non-linear interactions between stimulus elements. An advantage of this approach is that it maintains a clear distinction between associative and representational processes, making investigation of their interactions potentially more tractable. On the other hand, the accessibility of detailed elemental models such as the present one is limited by the potential complexity of the computations required to reveal the model's behavior.

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. 3, 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.

Aydin, A., & Pearce, J. M. (1997). Some determinants of response summation. *Animal Learning & Behavior, 25*, 108-121.

Bellingham, W. P., Gillette-Bellingham, K., & Kehoe, E. J. (1985). Summation and configuration in patterning schedules with the rat and rabbit. *Animal Learning & Behavior, 13*, 152-164.

Blough, D. S. (1975). Steady state data and a quantitative model of operant generalization and discrimination. *Journal of Experimental Psychology: Animal Behavior Processes, 104*, 3-21.

Brandon, S. E., Vogel, E. H., & Wagner, A. R. (2000). A componential view of configural cues in generalization and discrimination in Pavlovian conditioning. *Behavioural Brain Research, 110*, 67-72.

Brandon, S. E., & Wagner, A. R. (1998). Occasion setting: Influences of conditioned emotional responses and configural cues. *Schmajuk, Nestor A, (Ed); Holland*, DC, US: American Psychological Association.

Brimer, C. J. (1970). Disinhibition of an operant response. *Learning and Motivation, 1*, 346-371.

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

Bush, R. R., & Mosteller, F. (1951b). A model for stimulus generalization and discrimination. *Psychological Review, 58*, 413-423.

Dwyer, D. M., Mackintosh, N. J., & Boakes, R. A. (1998). Simultaneous activation of the representations of absent cues results in the formation of an excitatory association between them. *Journal of Experimental Psychology: Animal Behavior Processes, 24*, 163-171.

Estes, W. K. (1950). Towards a statistical theory of learning. *Psychological Review, 57*, 94-107.

Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science, 4*, 123-124.

Hall, G. (1991). Perceptual and associative learning. *New York, NY, US: Clarendon Press/Oxford University Press*.

Hall, G. (1996). Learning about associatively activated stimulus representations: Implications for acquired equivalence and perceptual learning. *Animal Learning & Behavior, 24*, 233-255.

Hall, G., & Honey, R. C. (1990). Context-specific conditioning in the conditioned-emotional-response procedure. *Journal of Experimental Psychology: Animal Behavior Processes, 16*(3), 271-278.

Harris, J. A., Jones, M. L., Bailey, G. K., & Westbrook, R. F. (2000). Contextual control over conditioned responding in an extinction paradigm. *Journal of Experimental Psychology: Animal Behavior Processes, 26*(2), 174-185.

Harris, J. A., Shand, F. L., Carroll, L. Q., & Westbrook, R. F. (2004). Persistence of preference for a flavor presented in simultaneous compound with sucrose. *Journal of Experimental Psychology: Animal Behavior Processes, 30*, 177-189.

Holland, P. C. (1990). Event representation in Pavlovian conditioning: Image and action. *Cognition, 37*, 105-131.

James, J. H., & Wagner, A. R. (1980). One-trial overshadowing: evidence of distributive processing. *Journal of Experimental Psychology: Animal Behavior Processes, 6*, 188-205.

Kamin, L. J. (1968). "Attention-like" processes in classical conditioning. In M. R. Jones (Ed.), *Miami symposium on the prediction of behavior: aversive stimulation* (pp. 9-31). Miami: Miami University Press.

Kehoe, E. J. (1982). Overshadowing and summation in compound stimulus conditioning of the rabbit's nictitating membrane response. *Journal of Experimental Psychology: Animal Behavior Processes, 8*, 313-328.

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.

Kehoe, E. J., Horne, P. S., Macrae, M., & Horne, S. J. (1993). Real-time processing of serial stimuli in classical conditioning of the rabbit's nictitating membrane response. *Journal of Experimental Psychology: Animal Behavior Processes, 19*, 265-283.

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

Lattal, K., & Nakajima, S. (1998). Overexpectation in appetitive Pavlovian and instrumental conditioning. *Animal Learning & Behavior, 26*(3), 351-360.

Lovibond, P. F., Preston, G. C., & Mackintosh, N. J. (1984). Context specificity of conditioning, extinction, and latent inhibition. *Journal of Experimental Psychology: Animal Behavior Processes, 10*, 360-375.

Lubow, R. E. (1973). Latent inhibition. *Psychological Bulletin, 79*, 398-407.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review, 82*, 276-298.

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

Mackintosh, N. J., & Bennett, C. H. (1998). Perceptual learning in animals and humans. In *Sabourin, Michel (Ed); Craik, Fergus (Ed); et al. (1998). Advances in psychological science, Vol. 2: Biological and cognitive aspects.* (pp. 317-333). Hove, England: Psychology Press/Erlbaum (UK) Taylor & Francis.

Mackintosh, N. J., & Reese, B. (1979). One-trial overshadowing. *Quarterly Journal of Experimental Psychology, 31*, 519-526.

McLaren, I. P., Bennett, C., Plaisted, K., Aitken, M., & Mackintosh, N. J. (1994). Latent inhibition, context specificity, and context familiarity. *Quarterly Journal of Experimental Psychology, 47b*, 387-400.

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). New York, NY, US: Clarendon Press/Oxford University 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., & Mackintosh, N. J. (2002). Associative learning and elemental representation: II. Generalization and discrimination. *Animal Learning & Behavior, 30*, 177-200.

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.

Pavlov, I. P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex.* (G. V. Anrep, Trans.). New York: Dover.

Pearce, J. M. (1987). A model for stimulus generalization in 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. (2002). Evaluation and development of a connectionist theory of configural learning. *Animal Learning & Behavior, 30*, 73-95.

Pearce, J. M., Aydin, A., & Redhead, E. S. (1997). Configural analysis of summation in autoshaping. *Journal of Experimental Psychology: Animal Behavior Processes, 23*(1) Jan), 84-94.

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., & 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. (1991). Failure of excitatory conditioning to extinguish the influence of a conditioned inhibitor. *Journal of Experimental Psychology: Animal Behavior Processes, 17*(4) Oct), 519-529.

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 a unique stimulus interpretation of configural conditioning. *Journal of Comparative and 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. (1988). Behavioral studies of Pavlovian conditioning. *Annual Review of Neuroscience, 11*, 329-352.

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

Rescorla, R. A. (2000). Associative changes in excitors and inhibitors differ when they are conditioned in compound. *Journal of Experimental Psychology: Animal Behavior Processes, 26*, 428-438.

Rescorla, R. A. (2001). Unequal associative changes when excitors and neutral stimuli are conditioned in compound. *Quarterly Journal of Experimental Psychology, 54B*, 53-68.

Rescorla, R. A. (2002a). Comparison of the rates of associative change during acquisition and extinction. *Journal of Experimental Psychology: Animal Behavior Processes, 28*, 406-415.

Rescorla, R. A. (2002b). Effect of following an excitatory-inhibitory compound with an intermediate reinforcer. *Journal of Experimental Psychology: Animal Behavior Processes, 28*, 163-174.

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

Rescorla, R. A., Grau, J. W., & Durlach, P. J. (1985). Analysis of the unique cue in configural discriminations. *Journal of Experimental Psychology: Animal Behavior Processes, 11*, 356-366.

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.

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

Spence, K. W. (1952). The nature of the response in discrimination learning in animals. *Psychological Review, 59*, 89-93.

Stevens, S. S. (1962). The surprising simplicity of sensory metrics. *American Psychologist, 17*, 29-39.

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

Wagner, A. D. (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. 5-47). Hillsdale, NJ: Erlbaum.

Wagner, A. R. (2003). Context-sensitive elemental theory. *Quarterly Journal of Experimental Psychology, 56B*, 7-29.

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, USA: Lawrence Erlbaum Associates, Inc.

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

Westbrook, R. F., Jones, M. L., Bailey, G. K., & Harris, J. A. (2000). Contextual control over conditioned responding in a latent inhibition paradigm. *Journal of Experimental Psychology: Animal Behavior Processes, 26*, 157-173.

Whitlow, J. W., & Wagner, A. R. (1972). Negative Patterning in classical conditioning: summation of response tendencies to isolable and configural components. *Psychonomic Science, 27*(5), 299-301.

Wilson, P. N., & Pearce, J. M. (1992). A configural analysis for feature-negative discrimination learning. *Journal of Experimental Psychology: Animal Behavior Processes, 18*(3), 265-272.

Notes

[1] This is because the model assumes that any stimulus occupies approximately half the available representational space. So two distinct stimuli, each of which independently activates half the total populations of elements, will share 25% percent of the total representational space, constituting 50% of their number of elements.

[2] I have confirmed this conclusion by conducting computer simulations of the two discriminations based on the McLaren and Mackintosh model.

[3] An attractive, and only marginally more complex, alternative describes element activation strength by a power function. In this scheme, the activation weight of an element is not linearly related to the physical intensity (I) of that feature, but is compressed as approximated by a power function with an exponent less than 1 (e.g., $\omega = I^{½}$). If the activation weight of an element were to correspond to the perceived intensity of that feature in the stimulus, then framing that relation in a power law is neither new nor controversial, it complies with long-standing psychophysical evidence concerning the relationship between the physical and perceived magnitudes of sensory events (Stevens, 1962). In this case, at the input level the physical intensity of a common element in a compound would still be the sum of its intensities in the two stimuli, but the activation weight of the corresponding element would be less than the sum of its weights in the single stimuli.

[4] It is worth noting that the associative strength that accrues to common elements under these circumstances is also limited in the amount of summation it can support. In the extreme case, where two identical CSs are conditioned (i.e, with 100% overlap), the input strength of all elements doubles in the compound (in this case, "the compound" being the same CS but with twice the intensity). However, since only half the elements can now enter the attention buffer, this will produce only a 50% increase in total activation weight. This is the same limit placed on summation when two CSs with no common elements are presented in compound.