The Mechanics of Associative Change

M.E. Le Pelley (mel22@hermes.cam.ac.uk)
I.P.L. McLaren (iplm2@cus.cam.ac.uk)
Department of Experimental Psychology; Downing Site
Cambridge CB2 3EB, England

Abstract

Rescorla (in press) investigated the change in associative strength undergone by cues A and B as a result of reinforcement or nonreinforcement of an AB compound. Many leading theories of associative learning predict that if A and B are equally salient then the associative change experienced by each should be the same regardless of their associative strength preceding AB trials. Rescorla explored this prediction for a compound composed of an excitatory A and an inhibitory B, using rats and pigeons as subjects. We repeated Rescorla's experiment using human subjects and a causal judgment task, and obtained diametrically opposite results to those of Rescorla's earlier study. The implications of this finding are discussed with reference to a number of influential theories of associative learning.

Introduction

It has long been recognised that stimuli presented in compound can, and will, interact and compete for associative strength. This is powerfully demonstrated in the phenomenon of blocking (Kamin, 1969). This refers to the finding that the gain in excitatory strength accruing to a conditioned stimulus (CS), B, following reinforcement (+) of an AB compound is much reduced if cue A has previously been trained as being a good predictor of that outcome (unconditioned stimulus, US). Learning does not simply progress with each cue independently. Instead the two cues seem to compete for a limited amount of associative strength. Such demonstrations of cue competition provided the motivation for the development of models that can deal with the issue of predictive redundancy in associative learning (e.g. Rescorla & Wagner, 1972; Mackintosh, 1975; Pearce, 1987).

A common feature of these models is the idea that the magnitude of associative change depends in some way on the discrepancy (or *error*) between the current associative strength of the presented cues and the strength which the outcome (unconditioned stimulus, US) following these cues can support. Consider, for instance, the Rescorla-Wagner (1972) model (R-W), perhaps the most influential of all of these "error-correcting" theories:

$$\Delta V_A = \alpha_A \beta_{US} (\lambda_{US} - \Sigma V) \qquad (1)$$

where ΔV_A is the change in associative strength of cue A, α_A and β_{US} are rate parameters relating to the salience of cue A and the US respectively, λ_{US} is the asymptote of conditioning supportable by that US, and ΣV is the summed associative strength of all cues present on a trial. Hence R-W states that the error governing associative change for any cue on a trial is based on the combined associative strength of *all* cues present on that trial. This is essential to R-W's explanation of blocking. On A+ trials, V_A will increase, with A coming to predict the US. On AB+ trials, the error term for cue B (and also for A) will be (λ-{V_A+V_B}). But of course as a result of A+ training V_A will already be high, and so the error term will be correspondingly small, such that any increase in V_B will be only very small. Thus the R-W explanation of blocking crucially hinges on the idea that, when determining the associative change undergone by B, the associative strength of other cues present on the trial (A) is also considered.

This use of a common error term governing associative change for all stimuli on a trial has important consequences. Rescorla (in press) noted that, in the absence of additional assumptions, it predicts that equally salient stimuli presented together on a trial will undergo equal associative changes. This prediction holds true regardless of the associative history of the cues in question.

In a recent series of experiments, Rescorla (in press) investigated this prediction in rats and pigeons (using magazine approach conditioning and autoshaping procedures respectively). He looked at the particular instance of an AB compound composed of an excitatory A and an inhibitory B. Specifically, he was interested in the associative change undergone by A and B as a result of either reinforcement or nonreinforcement of the AB compound. If we assume that A and B are of equal salience (ensured by counterbalancing) then, as a result of using a common error term, R-W is constrained to predict that both A and B will show equal associative change following either AB+ or AB- trials. Consider, for example, the AB+ condition. If A and B are equally salient, then α_A=α_B. Since both are presented with the same US, β will also be equal when calculating ΔV_A and ΔV_B according to equation (1). And finally, the error term for both A and B will be (λ-{V_A+V_B}). Hence, given that all the terms in the calculation for ΔV_A and ΔV_B are identical, Rescorla-Wagner must predict that on AB+ trials, ΔV_A=ΔV_B. This prediction of equal associative *change* holds true despite the fact that A (an excitor) and B (an inhibitor) begin these trials with very different associative strengths ($V_A > 0$, $V_B < 0$).

The problem with investigations such as this is one of how to assess the magnitude of associative change for two stimuli that differ in their "baseline" associative strength. It would be unwise to make any strong assumptions with regard to mappings between associative strength and measurable performance, and yet without such mappings we cannot be sure that two equal-sized changes in performance at different points on the performance scale represent equal-sized changes in associative strength.

Condition	Stage 1		Stage 2		Test
CR	A+	C+	AB+		**AD**
	E+		CD?		**BC**
	BE-	DE-			
CNR	F+	H+	FG-		**FI**
	J+		HI?		**GH**
	GJ-	IJ-			
Fillers	KL+	MN+	K-	M-	
	OP+	Q-	Q-	V+	
	R-	S-	W+	X+	
	T-	U-			

Table 1. Experimental design.
+: outcome; -: no outcome; ?: exposure trial.

Rescorla suggested an elegant way of avoiding this problem, by comparing performance to A and B when they were embedded in compounds designed to ensure comparable overall levels of performance. We adopt this technique in our experimental design (Table 1). Consider the Compound Reinforcement (CR) condition. A and C are initially trained as equivalent excitors, while B and D are trained as equivalent inhibitors. So following Stage 1, compounds AD and BC should have equal strengths, as each contains one of two equal excitors and one of two equal inhibitors. We then reinforce the AB compound. If this results in equal changes to the associative strengths of A and B (as predicted by R-W), then the AD and BC compounds should remain equal after Stage 2 (as each starts at the same level and receives the same change). If instead the strength of A increases more than that of B, then responding to AD will be greater than to BC. Conversely, if the strength of the inhibitory B increases more than that of the excitatory A, then BC will give rise to more conditioned responding than AD. A similar argument can be applied to the Compound Nonreinforcement (CNR) condition: if FG- trials cause a greater decrement in V_F than V_G then we expect the FI compound to be rated lower than GH: if V_G decreases more than V_F we expect the opposite.

Using this kind of logic, the results of Rescorla's experiments indicated that reinforcement of AB led to a greater increase in the associative strength of the inhibitory B than the excitatory A, while nonreinforcement of AB led to greater associative loss in the excitatory A than the inhibitory B. Rescorla (2001) carried out a similar investigation, this time with an AB compound consisting of an excitatory A and a neutral B. Again, AB+ trials led to a greater increase in V_B than V_A, whereas nonreinforcement gave a greater decrement in V_A than V_B. This indicated that the previous result was not simply due to some special property of conditioned inhibitors. Instead it seems that initial associative strength is an important factor in determining the distribution of associative change among the elements of a compound. This, of course, runs contrary to the predictions of R-W.

This prediction of equal change does not apply only to elemental theories such as R-W. Consider Pearce's (1987) configural model. This model proposes that a compound stimulus is best viewed as a unitary event that is separate from its elements, but able to generalise to them. Whereas according to an elemental model, an AB compound is decomposed into separate A and B elements, in a configural model it is represented as a single "AB" configuration. Generalised responding to other stimuli (e.g. A alone or B alone) occurs to the extent that these stimuli are similar to previously experienced configurations.

So on AB+ trials in Stage 2 of the CR condition, a configural model such as Pearce's learns an association between an AB configural unit and the US. Assuming that A and B are equally salient, this excitatory learning will generalise equally to each of them (they each have the same degree of similarity to AB), and so the model is constrained to predict equal associative change for A and B as a result of AB+ trials. Again no reference is made to the associative history of the cues. A similar story applies to the CNR condition: any change in association from an FG configural unit to the US will generalise equally to F and G.

Dickinson, Shanks & Evenden (1984) noted many similarities between Pavlovian conditioning in animals and the acquisition of causal judgments in human subjects. However, Le Pelley & McLaren (in press) demonstrated that not all phenomena in the animal learning field have analogues in studies of human causal learning. Given the importance of the previously described findings in elucidating the mechanisms underlying associative change, we aimed to repeat Rescorla's (in press) experiment using a causal judgment task with human subjects. This is of particular interest to us as, if we were to replicate Rescorla's findings in human subjects, it would invalidate the APECS model of associative learning that we have developed in recent years (Le Pelley & McLaren, in press; Le Pelley, Cutler & McLaren, 2000; see also McLaren 1993, 1994).

APECS is a model of learning and memory, based on the popular backpropagation algorithm (Rumelhart, Hinton & Williams, 1986), but with a couple of important differences. Firstly, APECS employs configural representation. Thus each different mapping of input to output is represented by its own hidden unit, which could equally well be termed "configural units". Secondly, APECS uses adaptive generalisation coefficients to determine the amount of generalisation between similar input patterns. As a result, once the weights appropriate to a mapping have developed, the learning in those weights can be protected against interference. This is achieved by reducing the learning rate parameter for the configural unit carrying that mapping. The effect is to "freeze" the weights to and from a certain configural unit at the value they hold immediately following experience of that configuration. Crucially, this freezing of weights to and from a certain configural unit occurs only if that configural unit has a negative error value, i.e. *if it is part of a mapping that predicts an incorrect outcome for the current input*. APECS has different learning rate parameters for input–hidden and bias–hidden connections. The former are frozen to prevent interference; the latter remain high. Hence extinction (suppression of inappropriate responses) is achieved by an increase in the negative bias on the hidden unit carrying the inappropriate mapping, rather than by reduction of weights (which would cause the original mapping to be lost from the network). Given appropriate input cues, the negative bias on the hidden unit can be overcome and the original mapping retrieved. So bias acts to change the retrievability of previously learnt mappings, such that

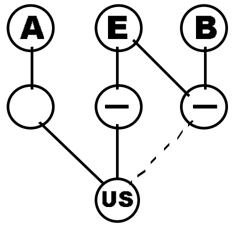

Figure 1. Associations developed following Stage 1, according to APECS. Excitatory connections are shown by solid lines, inhibitory associations by dotted lines. Negative bias on hidden units is indicated by a minus sign.

APECS addresses both learning (in formation of weights) and memory (in changes of retrievability).

Consider the processes at play in the CR condition, according to APECS. During Stage 1, the network will learn an excitatory connection from a representation of A to a configural unit, and from the configural unit to the output. Hence this configural unit comes to represent the A+ mapping. A different configural unit will be recruited to carry the E+ mapping. On BE- trials, presentation of E will cause positive activation to flow to the output via this E+ excitatory pathway. This positive activation is inappropriate on a trial on which the US is not presented: as a result the output unit will take on a negative error. This negative error will be propagated back along the excitatory connection to the E+ configural unit. This configural unit will therefore take on a negative error, as it is part of a mapping predicting an inappropriate outcome for the current input. This negative error means that, on BE- trials, the weights to and from the E+ configural unit will be frozen, as stated above. Instead the E+ configural unit will take on a negative bias to reduce expression of this excitatory mapping on these nonreinforced trials. In addition, excitatory associations will develop from B and E to a new configural unit, representing the BE configuration. This unit will develop an inhibitory association to the output in order to further counter any positive activation flowing to the output via the E+ hidden unit. By a similar argument, this BE- configural unit will develop negative bias on E+ trials. Hence following Stage 1, the situation for cues A, B and E is as shown in Figure 1 (this also applies to cues C, D, F, G, H and I and J, all of which have an equivalent partner in A, B or E following Stage 1).

What now happens on AB+ trials? The US will receive some positive activation via the A+ mapping learnt in Stage 1. However, it will also receive some negative activation via the BE- mapping (which can never be *totally* suppressed by development of negative bias). As a result the US will not be perfectly predicted on these trials, and yet is presented. Therefore the output unit will have a positive error. How can this error be reduced? Well, the positive error on the US unit will be propagated back to the BE- configural unit. But it is propagated along a negative connection, and so the BE- unit will take on a negative error (again, it is part of a mapping predicting an incorrect output on this trial). Thus weights to and from this unit are frozen. Extra negative bias can still be applied to the unit to reduce the negative activation flowing to the output, though, and this will help to reduce the output error. The positive output error will also be propagated back to the A+ configural unit along the positive connection. Thus the A+ configural unit will have a positive error. Both its weights and its bias are therefore free to change: the connections from A input to A+ configural, and from A+ configural to output, will increase. This too will reduce the output error.

In our previous expositions of APECS, we have always made the assumption that changes in weights occur faster than changes in bias. Thus we assume that changes due to learning take place faster than changes in memory, i.e. that learning represents rapid acquisition, and memory represents a more gradual decline in retrievability: this seems reasonable. We saw above that on AB+ trials, the weights of the A+ mapping are free to increase, whereas only the bias of the BE- mapping may change to reduce the effective strength of the inhibitory mapping. Therefore APECS is constrained to predict that, on these AB+ trials, the associative strength of the excitatory A will increase more than that of the inhibitory B. This is of course opposite to Rescorla's result.

A similar argument holds for the CNR condition. On FG- trials, the F+ hidden unit will have negative error (so only its bias may change), while the JG- unit has positive error (so that its weights and bias may both change). In this case APECS must predict that, on FG- trials, the associative strength of the inhibitory G will decrease more than that of the excitatory F, again opposite to Rescorla's result.

In summary, these results follow from the idea of adaptive generalisation. AB+ training generalises more to A+ than it does to BE- because AB and A predict the same outcome. Similarly FG- learning generalises more to GJ- than to F+ as FG and GJ predict the same outcome (no US).

Rescorla noted a problem with his paradigm that could cast doubt on the results obtained. AB compound presentation may result not only in development of A–US and B–US associations, but also in development of within-compound A–B associations. Consideration of these A–B associations complicates any inference of unequal associative change drawn from the results. Suppose A and B undergo equal changes as a result of AB+ trials. The formation of an A–B association might be expected to enhance responding to the inhibitory B (as it has been paired with an excitor), and reduce responding to the excitatory A (as it has been paired with an inhibitor). Thus even if the change in the A–US and B–US associations were equal, one would expect that AB+ trials would augment responding to B more than to A. Similarly, nonreinforcement of the AB compound might result in equal A–US and B–US decrements, but responding to A may fall further as it forms an association to the inhibitory B.

Rescorla controlled for the effect of within-compound associations in his Experiments 5 and 6. His findings were unchanged: Stage 2 AB+ trials gave a greater change in B than A, and *vice versa* for Stage 2 AB- trials. We were also careful to control for the effect of within-compound associations. In Stage 2, in addition to AB+ trials subjects also experienced CD "exposure trials". On these trials subjects saw cues C and D paired, but were not told whether or not the outcome occurred. Hence on these trials within-compound C–D–associations would form while C–US and D–US associations remain unchanged. The effect of A–B association formation would thus be matched by development of C–D associations. Therefore any difference between

AD and BC following Stage 2 could only be due to unequal changes in A–US and B–US associations on AB+ trials. The same holds true for the CNR condition.

Our investigation used an allergy prediction paradigm with human subjects. This paradigm has been used successfully in several studies of human causal learning (e.g. Dickinson & Burke, 1996; Le Pelley, Cutler & McLaren, 2000). Participants play the role of a food allergist judging the likelihood that various foods will cause an allergic reaction in a hypothetical patient. The foods, then, constitute the cues; the allergic reaction is the US. Following training, subjects rated how strongly certain individual foods, and compounds of two foods, predicted the occurrence of an allergic reaction. These ratings were taken as our measure of the strength of conditioning. This was a within-subjects experiment: subjects experienced all the different contingencies concurrently.

The Filler trials were included to ensure equal numbers of positive and negative trial types in each stage. In addition, they increase the number of different trial types seen by subjects, again following Dickinson & Burke and Le Pelley et al. This creates a large memory load, hopefully preventing subjects from basing their ratings on inferences made from explicit episodic memories of the various trial types. Instead subjects should have to rely on associative processes to provide an "automatic" measure of the causal efficacy for each cue. Using a large number of trial types makes us more confident that it is indeed associative, rather than cognitive, processes being tapped in our study.

Method

Participants Twenty members of Cambridge University (10 female, 10 male; age 19-49) took part in the experiment.

Procedure At the start of the experiment each subject was given a sheet of instructions presenting the "allergy prediction" cover story for the experiment. They were told that in the first block they would arrange for Mr. X to eat different meals on each day, and would monitor whether he had an allergic reaction or not as a result. In relation to the exposure trials (that do not bear on the issue at hand in this paper), subjects were told that occasionally the results of eating the foods had been lost. On these trials they would know the foods eaten in the meal, but not the result of eating those foods. They were also told that later on they would be asked to rate some of the foods according to how strongly they predicted allergic reactions.

On each conditioning trial, the words "Meal [meal number] contains the following foods:" followed by the two foods appeared on the screen. Subjects were then asked to predict whether or not eating the foods would cause Mr. X to have an allergic reaction, using the "x" and "." keys (counterbalanced). The screen then cleared, and immediate feedback was provided. On positive trials the message "ALLERGIC REACTION!" appeared on the screen; on negative trials the message "No Reaction" appeared. If an incorrect prediction was made, the computer beeped. On the exposure trials of Stage 2, the same message appeared, but now subjects were cued to enter the initial two letters of each of the foods. This was to ensure that they paid attention to the pairings of foods when no allergy prediction was required. The 24 foods used were randomly assigned to the letters A to X in the experimental design for each subject.

As shown in Table 1, there were 16 trial types in Stage 1, and 8 in Stage 2. Stages 1 and 2 were split into 8 sub-blocks, with each trial type appearing once in each sub-block (hence subjects saw each trial type 8 times). The order of trials within each sub-block was randomised, as was the order of presentation on the screen (first/second) within each compound pair.

After Stage 1, subjects were asked to rate their opinions of the effect of eating certain foods on a scale from -10 to +10. They were to use +10 if the food was very likely to cause an allergic reaction in Mr. X, -10 if the food was very likely to prevent the occurrence of allergic reactions which other foods were capable of causing, and 0 if eating the food had no effect on Mr. X (i.e. it neither caused nor prevented allergic reactions). For clarification, participants also had access to a card on which the instructions on how to use the rating scale were printed. Once a meal had been rated it disappeared from the screen and the next appeared: participants could not revise their opinions upon seeing later meals. Subjects were given a second rating test after Stage 2, when they were asked to rate meals containing either one food or two. Exactly the same test procedure was used.

Results and Discussion

The results of Test 1 indicated that we had been successful in generating conditioned excitation (to A and F, mean rating 8.63) and inhibition (to B and G, mean rating –8.1), as compared to Q (mean rating 0.6), which is never paired with the US and hence should remain neutral. Planned comparisons revealed that the average of A and F (which are equivalent) was significantly higher than Q, which was in turn significantly higher than B and G (which are also equivalent) [*F*(1,19)=65.9 and 93.2 respectively, *p*s<0.001].

Figure 2 shows the mean rating of the casual efficacy of each of the meals of interest as judged in the test following Stage 2. We see that the compound AD is rated higher than BC. This is confirmed statistically [*F*(1,19)=5.87, *p*<0.05]. This implies that reinforcing the AB compound led to a greater increase in the associative strength of the excitatory A than the inhibitory B. This is, of course, diametrically opposite to Rescorla's earlier findings with rats and pigeons.

In addition, we see that FI is rated significantly higher than GH [*F*(1,19)=4.43, *p*<0.05]. This implies that nonreinforcement of the FG compound led to a greater decrement in the associative strength of the inhibitory G than the excitatory F. Again, this is opposite to Rescorla's findings.

Like Rescorla's, our results suggest that the distribution of associative changes among the elements of a compound depends on the associative history of those elements. This asymmetry in associative change contradicts any model employing a common error term governing associative change

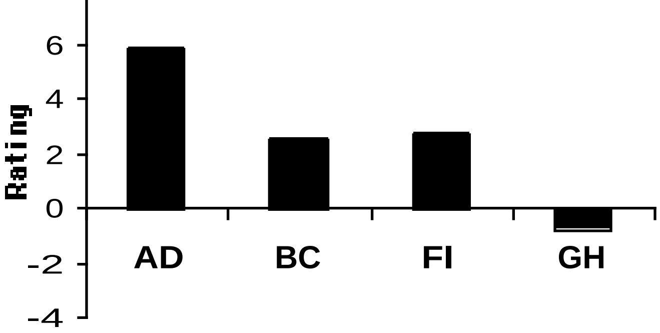

Figure 2. Mean ratings given to the cues of interest.

for all stimuli present, as these models are constrained to predict equal changes for these stimuli.

However, unlike Rescorla's results, we found that it is the cue whose associative strength is *less* discrepant from that supported by the outcome of the trial that undergoes the greater associative change. Our empirical findings agree with the predictions of the APECS model of learning and memory outlined earlier. This was confirmed by simulation. We performed 20 simulations with APECS, each representing a different subject. Each trial involved 1000 learning cycles. A hidden unit is defined as being "active" when it receives positive activation from the input layer. Thus if cue A is presented to the network, any hidden unit representing a configuration that includes cue A will be active. Activity extends into the period immediately following each trial, when no inputs are presented (again for 1000 learning cycles). The learning rate parameters for input–hidden and hidden–output units are both 0.8 when a hidden unit is active and has a positive error, and 0 when it is not. The parameter for bias–hidden changes is 0.3 when a hidden unit is active, 0 when it is not. The exact values of these parameters are unimportant: the pattern of results is robust under quite large variations in the values used. The results of the simulation are shown in Figure 3. As expected, we see that AD is rated higher than BC, and FI is rated higher than GH. Both differences are significant [$F(1,19)=20.2$ and 3.2 respectively, $ps<0.05$]. This is, of course, the pattern seen in our empirical data, and the opposite of Rescorla's results.

The theories explicitly considered thus far describe cue competition effects in terms of modifications in the effectiveness of the US. If the US is surprising (i.e. error is high) then it is able to support more learning than if it is already predicted (and therefore less surprising). The contribution of the CS to learning is assumed to be fixed, and is determined by its salience (α, or the parameter for learning weights in APECS). In addition to such US-centred views, there exist a number of influential theories of associative learning that instead ascribe cue competition to variations in the processing of the CS. For example, blocking of B on AB+ trials following A+ pretraining might be interpreted as reduced processing of B as a result of earlier learning about A's predictive power. Typically, these CS-processing models specify a role for attentional processes in determining the distribution of associative change undergone by the cues of a compound. The attention paid to a cue depends on the associative history of that cue, so perhaps these theories would be better-suited to explaining our results.

Mackintosh (1975) proposed just such a model of selective attention. This theory states that good predictors of an outcome will retain a higher salience (i.e. will receive greater attention) than poorer predictors. The calculation determining the attention to be paid to stimulus A relies on a comparison of the predictive power of A *for the outcome occurring on that trial* with the predictive power of all other presented cues for that same outcome. This calculation is carried out after every trial. If cue A is a better predictor of the outcome of that trial than any other cue present, its salience increases for the next trial, and *vice versa*. Hence this model proposes that CSs followed by their expected outcomes (be this reinforcement or nonreinforcement) garner greater salience. CSs followed by surprising events (again, be this reinforcement or omission of reinforcement) lose salience.

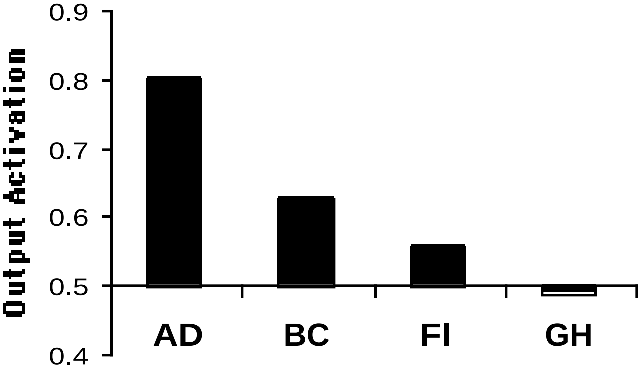

Figure 3. Simulation of the data using APECS.

According to the Mackintosh theory, learning about each element of an AB compound is governed by the discrepancy between λ and the *individual strength of that element* (rather than the discrepancy between λ and the summed strengths of A and B), modulated by the attention it receives. Thus:

$$\Delta V_A = \alpha_A \beta_{US} \left(\lambda_{US} - V_A \right) \qquad (2)$$

where α_A represents the attention paid to cue A.

Consider the CR Condition of our experiment. During Stage 1, A is consistently followed by the US, and B is consistently followed by no US. Thus both will begin Stage 2 with fairly high salience, as they are both good predictors of their respective "outcomes" (which for B is actually nonreinforcement). On Stage 2 AB+ trials, however, A is a better predictor of the outcome (reinforcement) than B, which predicts nonreinforcement. Hence attention to A will remain high, while that for B will be reduced rapidly: increments in V_A will remain relatively high over Stage 2 trials, while increments in V_B will become progressively smaller. As a result, Mackintosh (1975) is able to predict that over all Stage 2 trials, the increment in V_A will be greater than that for V_B. This is, of course, exactly the pattern seen in our empirical data. A similar story holds for the CNR contingency – on Stage 2 FG- trials, the inhibitory G is a better predictor of nonreinforcement than the excitatory F. Hence attention to G will remain high over Stage 2, while attention to F will fall. So overall we might expect a greater decrement in responding to G than to F.

In general, then, and in agreement with our data, Mackintosh (1975) is able to predict that the stimulus whose associative strength is less discrepant from the outcome of the trial (i.e. the better predictor of the outcome) will show the greater change on compound training.

Intriguingly Mackintosh's (1975) can also explain Rescorla's empirical data, which are diametrically opposed to our own, by appealing to the notion of overtraining (Mackintosh, personal communication). If we train subjects on Stage 1 until the associative strengths of excitors and inhibitors closely approach their asymptotic values, then the predictions made by the theory change dramatically.

As a result of this overtraining, A and B will also have very high salience (near asymptote) at the start of Stage 2.

On the initial AB+ trial, then, both will be well processed (as the calculation to update the salience of a cue is performed *after* each trial). Given that the error term governing associative change for a cue involves only the current associative strength of that cue, rather than the summed strength, the stimulus whose associative strength is more discrepant from that supportable by the outcome of the trial will undergo greater change. In other words, on the initial AB+ trial, it will actually be the *poorer* predictor of the trial's outcome (B) that undergoes the greatest associative change, as this cue will have the greater error term. Notably, if A's associative strength is near asymptote (λ), its error term according to equation (2) will be near zero. Given that it is error that drives changes in associative strength, this means that any change in V_A will be only very slight. Of course the modulation of attention discussed earlier will still occur. Thus following this initial trial, attention to A (a good predictor of the outcome) remains high, while that for B (a poor predictor of the outcome) will be reduced. So subsequent changes in V_B will become increasingly smaller. However, given that V_A was already near asymptote at the start of Stage 2, it will undergo little further increase over Stage 2 trials. In other words, the effect of Stage 1 training outweighs any influence of attentional modulation in Stage 2. In fact, the effect of attentional modulation may be reduced even further in the case of an overtrained contingency if we follow Sutherland & Mackintosh's (1971, p. 491) suggestion that the high attentional strengths developed as a result of overtraining are "sticky": a high α value is reduced more slowly than an intermediate value. As such the high value of α_B will persist over several AB+ trials. This, combined with B's high error value on these trials, will result in large increments in V_B over several trials before more significant reductions in α_B start to take their toll on the size of the increments. Thus by appealing to overtraining in Stage 1, Mackintosh (1975) is able to predicts Rescorla's finding that V_B increases more than V_A as a result of AB+ trials.

On this analysis, then, the difference between Rescorla's experiment and our own is that in the former, Stage 1 conditioning led to near-asymptotic associative strengths such that any effect of selective attention in Stage 2 was outweighed. The notion of sticky α values developed as a result of overtraining will further reduce the influence of attentional processes. In our experiment, however, we must assume that Stage 1 training did not approach asymptotic levels, such that both cues in the Stage 2 compound were free to undergo associative change as dictated by their salience.

Conclusion

In common with Rescorla's earlier experiments, our results indicate that a cue's associative history is important when determining the magnitude of its associative change. Unlike Rescorla's experiments, however, our data indicate that it is the better predictor of an outcome that undergoes the greater associative change on compound conditioning. This finding may reflect different rules governing the distribution of associative change among elements of a compound in humans and animals, or may simply be a result of different levels of initial training in the human and animal studies. Given only the results of the current experiment we cannot choose between a "US processing" model of learning and memory employing adaptive generalisation with configural representation (APECS), and a model of selective attention in which CSs compete for attention (Mackintosh, 1975). However, taken in conjunction with several other findings from this laboratory (Le Pelley, Cutler & McLaren, 2000, Le Pelley & McLaren, in press; Le Pelley & McLaren, this issue), we believe that the results of all the human data may be better explained by APECS.

References

Dickinson, A., & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgments. *Quarterly Journal of Experimental Psychology, 49B,* 60-80.

Dickinson, A., Shanks, D. R., & Evenden, J. (1984). Judgment of act-outcome contingency: The role of selective attribution. *Quarterly Journal of Experimental Psychology, 36A,* 29-50.

Kamin, L.J. (1969). Selective association and conditioning. In N.J. Mackintosh & W.K. Honig (Eds.), *Fundamental issues in associative learning* (pp. 42-64). Halifax: Dalhousie University Press.

Le Pelley, M.E., Cutler, D.L., & McLaren, I.P.L (2000). Retrospective effects in human causality judgment. *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 782-787). Hillsdale, NJ: Lawrence Erlbaum Associates.

Le Pelley, M.E., & McLaren, I.P.L. (this issue). Representation and generalization in associative systems.

Le Pelley, M.E., & McLaren, I.P.L. (in press). Retrospective revaluation in humans: Learning or memory? *Quarterly Journal of Experimental Psychology B.*

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

McLaren, I. P. L. (1993). APECS: A solution to the sequential learning problem. *Proceedings of the XVth Annual Convention of the Cognitive Science Society* (pp. 717-722). Hillsdale, NJ: Lawrence Erlbaum Associates.

McLaren, I. P. L. (1994). Representation development in associative systems. In J.A. Hogan & J.J. Bolhuis (Eds.), *Causal mechanisms of behavioural development* (pp. 377-402). Cambridge: Cambridge University Press.

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

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. (in press). Associative changes in excitors and inhibitors differ when they are conditioned in compound. *Journal of Experimental Psychology: Animal Behaviour Processes.*

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.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D.E. Rumelhart, J.L. McLelland & the PDP Research Group (Eds.), *Parallel Distributed Processing* (Vol. 1, pp. 318-362). Cambridge, MA: MIT Press.

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