Learning & Behavior
2005, 33 (2), 147-159

Simulations of a modified SOP model applied to retrospective revaluation of human causal learning

MICHAEL R. F. AITKEN and ANTHONY DICKINSON
University of Cambridge, Cambridge, England

Dickinson and Burke (1996) proposed a modified version of Wagner's (1981) SOP associative theory to explain retrospective revaluation of human causal judgments. In this modified SOP (MSOP), excitatory learning occurs when cue and outcome representations are either both directly activated or both associatively activated. By contrast, inhibitory learning occurs when one representation is directly activated while the other is associatively activated. Finite node simulations of MSOP yielded simple acquisition, overshadowing, blocking, and inhibitory learning under forward contingencies. Importantly, retrospective revaluation was predicted in the form of unovershadowing and backward inhibitory learning. However, MSOP did not yield backward blocking. These predictions are evaluated against the relevant empirical evidence and contrasted with the predictions of other associative theories that have been applied to retrospective revaluation of human causal and predictive learning.

When Dickinson, Shanks, and Evenden (1984) first suggested that common associative learning processes mediate the acquisition of causal judgments and Pavlovian conditioning, they did so for two reasons. First, they demonstrated that the profile of causal judgments that they observed under variations in the contingency between a putative causal event and the outcome were simulated by associative theories of Pavlovian conditioning. Second, they found that the acquisition of causal judgments was subject to a form of selective learning, blocking, that had previously been a touchstone for developing these associative theories (Kamin, 1969).

Blocking is theoretically important as a demonstration that associative learning is driven by surprising or unexpected outcomes. In a typical blocking procedure, a compound of a treatment cue T and a target cue X is paired with an outcome (TX^+). The amount learned about the target cue X is reduced if the outcome is predicted by the treatment cue T as a result of prior training with the outcome (T^+). Under this blocking (T^+TX^+) contingency, learning about the target cue X is said to be blocked by the treatment cue T (Kamin, 1969). In their human causal learning task, Dickinson et al. (1984) found that following training on a probabilistic version of the blocking (T^+TX^+) contingency, causal judgments for the target event X were reduced or blocked relative to those for event X from a control (TX^+) contingency in which the treatment cue T was not pretrained.

Although a number of associative theories have been applied to human causal learning over the intervening 20 years, we shall focus on Wagner's (1981) SOP model, which was originally formulated as a real-time, process-based theory of Pavlovian conditioning and first applied to causal learning by Dickinson and Burke (1996). In SOP, stimuli are represented by nodes in associative memory, each composed of a number of elements. Wagner proposed that these elements could have three different activation states: an inactive state, I, and two active states, A1 and A2. Unpredicted presentation of a cue or outcome activates a proportion of the elements in the corresponding node from I into A1. Over time, the A1 state of these elements decays into A2, and then back into I. Importantly, activating a node by an associative connection bypasses A1 and leads to a direct transition from I to A2. This property of the model is critical because it allows SOP to explain a variety of learning effects.

Wagner (1981) identified two forms of learning that occur whenever the elements of nodes are concurrently active, which are represented in the top row of Table 1. First, there is an increment in the strength of an excitatory association between nodes representing these stimuli to the extent that the elements of the nodes are concurrently in A1. This excitatory learning occurs when a novel cue is paired with an unpredicted outcome because at the outset of training, the presentation of the cue and outcome excites the elements of their respective nodes into A1. Blocking also follows, at least in part, from the conditions for excitatory learning. Pretraining of the treatment cue T enables the presentation of this cue to activate some of the outcome elements into A2 prior to the presentation of the outcome, so that when the out-

This research was supported by Biotechnology and Biological Sciences Research Council of the United Kingdom Grant S05642 and conducted within the Medical Research Council Centre for Behavioural and Clinical Neuroscience. Correspondence should be addressed to A. Dickinson, Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, England (e-mail: ad15@cam.ac.uk).

Table 1
The Modified SOP (MSOP) Model

	Outcome Element State	
Cue Element State	A1	A2
A1	Ex	In
A2	In	Ex

Note—Ex, excitatory learning; In, inhibitory learning.

come occurs during TX$^+$ compound training, fewer of its elements are available for activation from I to A1. Consequently, the reduced number of outcome elements concurrently in A1 with target cue X elements attenuates the amount of excitatory learning to cue X.

Furthermore, according to SOP, a second learning process also makes a contribution to blocking. The blocking (T$^+$TX$^+$) contingency ensures that cue X elements are in A1 concurrently with outcome elements in A2, the latter elements having been driven from I directly into A2 by the presentation of the pretrained treatment cue T. Within SOP, concurrent activation of cue elements in A1 and outcome elements in A2 leads to the strengthening of an inhibitory association between the cue and outcome nodes (see first row of Table 1).

An application of SOP to causal learning assumes that judgments of the causal status of a cue are determined by the overall associative strength of the cue when aggregated across its excitatory and inhibitory associations with the outcome. If the aggregate associative strength is excitatory, then the cue is judged to cause the outcome, with the effectiveness of this generative causation being determined by the aggregate excitatory associative strength. Therefore, blocking reflects a reduction in this aggregate excitatory strength produced by both a reduction in the excitatory association and an increment in the inhibitory association between cue and outcome representations.

The application of the model also assumes that cue–outcome associations that are on aggregate inhibitory are manifest as judgments that the cue is a preventative cause that tends to prevent the outcome from occurring. According to SOP, the canonical contingency producing a preventative cause is one in which the treatment cue T alone is paired with the outcome (T$^+$), whereas a compound of the treatment cue T and the target cues X is presented without the outcome (TX$^-$). Under this inhibitory (T$^+$TX$^-$) contingency, the presentation of treatment cue T on the TX$^-$ episodes associatively activates the outcome elements in A2 so that there is A1 activation of target cue X elements concurrently with A2 activation of the outcome elements, a condition that establishes a pure inhibitory association (see first row of Table 1). Ever since Pavlov's (1927) original characterization of conditioning, it has been known that the inhibitory (T$^+$TX$^-$) contingency establishes cue X as a conditioned inhibitor, and within a causal learning scenario this contingency leads to cue X being rated as a preventative cause (e.g., Aitken, Larkin, & Dickinson, 2000; Chapman, 1991).

In spite of the success of associative theories in explaining basic effects in causal learning (see De Houwer & Beckers, 2002b, for recent review), such theories impose a major limitation on the integration of information relevant to causal judgments. This limitation is most clearly illustrated by event contingencies that require the retrospective revaluation of causal judgments. In a follow-up to the Dickinson et al. (1984) blocking experiment, Shanks (1985) investigated the effect of reversing the two stages of the standard forward blocking (T$^+$TX$^+$) contingency to generate a backward blocking (TX$^+$T$^+$) contingency. Theoretically, this reversal is interesting because, from a normative standpoint, it should have no effect on judgments about the target cue X. At the time of judgment, the participants have been exposed to exactly the same information about the pairings of the target and treatment cues with the outcome in the forward and backward contingencies. Consequently, if the forward contingency yields a lower causal judgment relative to the target cue from the control (TX$^+$) contingency, so should the backward contingency.

However, backward blocking is problematic for associative analyses of causal learning. Unlike forward blocking, the backward form requires the retrospective use of information acquired during the training of the treatment cue T alone to reevaluate the status of the target cue X. To illustrate why this retrospective revaluation presents a problem for associative theory, we consider the application of SOP to backward blocking (TX$^+$T$^+$). A simplified application of the model is illustrated in the top two rows of Table 2, which display the initial activation states engaged on the first trial or episode of the second stage of the retrospective revaluation contingencies but omit any within-trial state transitions. According to SOP, four excitatory associations should be formed during the first compound TX$^+$ training stage: two associations with the outcome representation, one from the treatment cue T and the other from the target cue X, and two within-compound associations between the cues T and X themselves. Given these associations, the presentation of the treatment cue T during the second stage has two consequences. The first is the activation of some of the elements of the outcome node into A2. The subsequent presentation of the outcome then activates a proportion of the remaining outcome elements into A1 so that the outcome node has a mixed activation state with some elements in A1 and some in A2.

The second effect of presenting the treatment cue T is to activate elements of the target cue X node into A2 via the within-compound association. According to SOP, however, the conjoint activation of the target cue X elements in A2 and outcome elements into the mixed A1 and A2 states has no impact on the associative strengths of cue X. Standard SOP (Wagner, 1981) assumes that the associative strengths of a cue can be changed only when at least some of its elements are in A1, and it is for this reason that any form of retrospective revaluation lies outside the scope of the model. In response to this limi-

Table 2
Application of the Modified SOP (MSOP) Model to Retrospective Revaluation

		Element States During Stage 2		
Stage 1	Stage 2	Cue	Outcome	Learning
		Backward Blocking		
TX^+	T^+	A1	A1 & A2	Ex & In
	\			
	x	A2	A1 & A2	Ex & In
		Unovershadowing		
TX^+	T^-	A1	A2	In
	\			
	x	A2	A2	Ex
		Backward Inhibition		
TX^-	T^+	A1	A1	Ex
	\			
	x	A2	A1	In

Note—T, treatment cue; X, target cue; x, representation of target cue X retrieved via a within-compound association (\) with the treatment cue T; $^+$, outcome; $^-$, no outcome; Ex, excitatory learning; In, inhibitory learning.

tation, Dickinson and Burke (1996) suggested that the model might be modified to allow learning to occur when the cue elements are in A2. The nature of this learning is illustrated in Table 1. Specifically, excitatory associations are incremented whenever the elements of two nodes are in the same state, be it A1 and A1 or A2 and A2, whereas inhibitory learning occurs when the two nodes are in different activation states.

As Table 2 shows, the application of this modified SOP (MSOP) yields an ambiguous prediction for retrospective revaluation under the backward blocking (TX^+T^+) contingency. On the initial episode or trial of the second T^+ stage, the pairing of the cue X elements in A2 with outcome elements in both A2 and A1 will lead to increments in both the excitatory and inhibitory associative strengths of cue X, and only if the latter influence predominates will backward blocking be observed. In fact, the empirical evidence for backward blocking of causal and contingency judgments is mixed. Whereas Shanks (1985) and others (e.g., De Houwer, Beckers, & Glautier, 2002; Le Pelley & McLaren, 2001; Wasserman & Berglan, 1998) have reported backward blocking in a causal scenario, Larkin, Aitken, and Dickinson (1998) consistently failed to find any evidence for this form of retrospective revaluation.

MSOP does, however, unambiguously predict another form of retrospective revaluation: Presenting the treatment cue T alone after TX^+ compound training (TX^+T^-) augments causal judgments for the target cue X (Larkin et al., 1998; Wasserman & Berglan, 1998). We refer to this effect as unovershadowing (recovery or release from overshadowing) because it is presumed to reflect a reversal of the assumed overshadowing of learning about cue X by cue T during the first, TX^+ compound, training stage (Kamin, 1969).

According to MSOP, the associations established in the first compound stage of the unovershadowing (TX^+T^-) contingency are the same as those in the backward blocking contingency, so that the presentation of cue T alone in the second stage activates elements for both cue X and the outcome conjointly into A2 (Table 2). By contrast to backward blocking, however, the absence of the outcome ensures that none of its elements are in A1, so that the conjoint A2 A2 activation leads to pure excitatory learning about the target cue X on the T^- episodes. Consequently, the causal ratings for cue X should be enhanced relative to the target cue from the control (TX^+) contingency in which the treatment cue is not presented during the second stage.

Another form of retrospective revaluation predicted by MSOP is generated by reversing the forward inhibitory (T^+TX^-) contingency. Recall that according to standard SOP, this contingency endows the target cue X with inhibitory associative strength. MSOP makes exactly the same prediction for the backward inhibitory (TX^-T^+) contingency, in which the two stages of the inhibitory contingency are reversed. As Table 2 shows, the backward inhibitory (TX^-T^+) contingency ensures that when the treatment cue T is initially paired with the outcome, elements of target cue X are in A2 jointly with elements of the outcome in A1. Within MSOP, this activation pattern supports pure inhibitory learning, which should be manifest as the acquisition of preventative status by cue X. The acquisition of preventative judgments under a backward inhibitory contingency has been documented in a number of experiments (e.g., Chapman, 1991; Larkin et al., 1998; Williams & Docking, 1995).

In summary, MSOP provides a potential explanation of the pattern of generative and preventative causal judgments acquired under a variety of contingencies, including retrospective ones. However, the derivation of these predictions has been informal and descriptive. Moreover, the predictions are based entirely on the initial distribution of activation states of the elements on the first episode with the treatment cue T alone and ignore the changes in state distributions that occur both within each trial and across trials. Consequently, given the dynamic nature of

MSOP, which is inherited from its parent model, we cannot be confident of the predictions of the model without simulation. Nor can we assess the robustness of these predictions across parameter variations on the basis of a descriptive analysis alone. The purpose of the present study was to explore the predictions of a simulation of MSOP. In doing so, we focused on the predictions of the model under six contingencies using the training conditions employed by Larkin et al. (1998): overshadowing (TX^+) and unovershadowing (TX^+T^-); forward (T^+TX^+) and backward (TX^+T^+) blocking; and forward (T^+TX^-) and backward (TX^-T^+) inhibition.

IMPLEMENTATION OF MSOP

The implementation of MSOP was written in C++, compiled using Microsoft Visual C++ 6.0. The model contains two states, A1 and A2, corresponding to the two activation states within SOP, and a set of representational elements. A1 and A2 consist of sets of locations each of which may contain a single representational element. No element may occupy a slot in both A1 and A2 simultaneously; and any element that is not contained in either state is in the inactive (I) state. Whenever an element moves into one of the two states, it displaces the contents of a single, randomly selected location within the state. Any element displaced from A2 enters I, whereas any element displaced from A1 immediately enters into A2, thereby displacing the contents of a single, randomly selected location. Each stimulus is represented by a node that contains a set of elements, and modifiable connection strengths exist between all stimulus nodes.

In order to approximate the real-time nature of SOP, we implemented MSOP by state changes across a series of timeslices with the following events within a timeslice being treated as occurring simultaneously (although, by necessity, calculated and performed sequentially):

1. Inactive elements of all stimuli presented during the timeslice have a nonzero probability ($P_{\mathrm{I} \rightarrow \mathrm{A1}}$) of moving into A1. Each transition occurs independently, and if two elements move into A1 at this stage, there is a chance that one of these elements will immediately displace the other. In order to prevent systematic effects of such displacements, whenever two or more stimuli are presented simultaneously, a single element from each stimulus in turn is given an opportunity for promotion from I to A1. Allowing a chance of promotion into A1 for all elements of stimulus Y, followed by those of stimulus Z, would lead to a higher proportion of promoted elements of Y being instantaneously displaced into A2.

2. The degree of conjoint activity between the elements of stimulus pairs is then used to calculate a matrix of associative strength changes between all stimulus nodes, according to the following equation:

$$\delta V_{\mathrm{YZ}} = l\{\mathrm{Y_{A1}Z_{A1}} - \mathrm{Y_{A1}Z_{A2}}/r \\ -\rho(\mathrm{Y_{A2}Z_{A1}} - \mathrm{Y_{A2}Z_{A2}}/r)\}, \qquad (1)$$

where δV_{YZ} is the change in connection strength between nodes Y to Z; l is a general learning rate parameter; ρ is the ratio of the modified (A2-A1) and standard (A1-A1) leaning rates; r is the ratio of the size of the A2 and A1 states, and $\mathrm{Y_{A1}}$ is the proportion of elements of Stimulus Y in the A1 state (similarly $\mathrm{Y_{A2}}$, $\mathrm{Z_{A1}}$, etc.). In SOP (Wagner, 1981), separate learning rules are specified for changes in excitatory and inhibitory associative strength that are then combined into a net associative strength. Although the distinction between excitatory and inhibitory learning is important psychologically, it does not affect the computational implementation in the present context, and therefore we implemented learning in terms of the net associative strength for simplicity.

3. At the end of each timeslice, elements may move from I to A2. A1 activity for a stimulus that has a positive (excitatory) link to a second stimulus increases the probability of inactive elements of the second stimulus moving into A2. A1 activity in a stimulus that has a negative (inhibitory) link to a second stimulus decreases this probability:

$$if \sum V_{\mathrm{YZ}}\mathrm{Y_{A1}} > 0: \quad P_{\mathrm{Z:I} \rightarrow \mathrm{A2}} = \sum V_{\mathrm{YZ}}\mathrm{Y_{A1}}, \qquad (2)$$

$$if \sum V_{\mathrm{YZ}}\mathrm{Y_{A1}} < 0: \quad P_{\mathrm{Z:I} \rightarrow \mathrm{A2}} = 0, \qquad (3)$$

where V_{YZ} is the current associative strength of the link from stimulus Y to stimulus Z, summation is across all stimuli (other than Z), and $P_{\mathrm{Z:I} \rightarrow \mathrm{A2}}$ is the probability of any given inactive element of Z moving into A2 during the timeslice.

4. Between the end of each timeslice and the beginning of the next, the matrix of associative strengths is adjusted by addition of the δV matrix calculated within the timeslice. In addition, N_d insertions are made of null elements into A1, forcing displacement (decay) of elements from A1 to A2 and from A2 to I. Psychologically, we assume that these insertions are generated by background or contextual stimulation.

With ρ set to zero, the implemented model thus corresponds to the specification of SOP as originally suggested by Wagner (1981) except in one respect. This difference is manifest in Equation 2: A2 activity within other stimulus nodes has no impact on evoked A2 activity, whatever the associative strength between them.

In SOP, the degree to which one stimulus (X) may come to evoke an activation in another stimulus node (Z) might be due to not only to any direct association between the two, but also by both stimuli being associatively linked to a third stimulus (Y). By removing the ability of A2 activity in one node to evoke A2 activity in another, the activity evoked in Node Z by presentation of X will be solely a result of the direct associative connection between them. This difference between SOP and MSOP will make little change to the logic of the models: In the original formulation, indeed, Wagner (1981) noted that we “may assume that the retrieval influence of A2 state elements was, in fact, negligible” (p. 23). Furthermore, any procedure that results in association between

X and Y, and also between Y and Z, will according to MSOP result in the opportunity for direct associations to be formed between X and Z, due to the new learning processes involving absent cues. In essence, while this change will alter the prediction of MSOP for such designs, this is precisely what the modification to learning rules in the model was designed to achieve.

The omission of A2-A2 retrieval also removes a possible instability in Dickinson and Burke's (1996) original descriptive specification of MSOP, in which mutual excitatory connections between a pair of nodes could lead to an unlimited increase in mutual excitatory learning if the retrieval influence of A2 activity exceeds the rate of decay A2 to I.

SIMULATIONS OF MSOP

In the simulations, cue–outcome pairings occurred in trials with each trial consisting of a series of timeslices. Cues were presented on Timeslices 1–10, and the outcome, if present, on Timeslices 11–20 of each trial. All associative strengths were zero at the start of training, and all elements were in I at the start of each trial. Each trial continued for as long as was required for all the elements of the stimulus nodes for each presented stimulus to return to the inactive state.

For all simulations, the size of A1 was 1,000 elements, and the ratio (r) of A2 size to A1 size was set to 5 (i.e., 5,000 elements in A2), reflecting the ratio used by Wagner (1981).

The simulations were applied to the various contingencies under the training conditions used by Larkin et al. (1998), which involved simultaneous compound presentations of treatment cue T and target cue X for six trials and three trials of training for the treatment cue T. On both types of trial, the cue(s) could either be presented alone (TX^-, T^-) or followed by the outcome (TX^+, T^+), depending upon the contingency.

First Treatment Cue Trial

By comparison with SOP, the innovative feature of MSOP is its capacity to explain retrospective revaluation. In the introduction, we presented a descriptive account of how the interaction between A1 and A2 processes generates retrospective revaluation on the first trial of the second stage when the treatment cue T is trained by itself. Figure 1 illustrates this account by simulating the temporal dynamics of A1 and A2 on this trial for the three retrospective contingencies: unovershadowing (TX^+T^-; left panel), backward blocking (TX^+T^+, middle panel), and backward inhibition (TX^-T^+; right panel). The parameters were $l = 0.01$; $\rho = 0.1$; $N_d = 50$; $P_{I \rightarrow A1} = .05$; and all stimulus nodes contained 400 elements ($N_s = 400$), with no element being in more than one stimulus. These choices were arbitrary, selected for illustrative purposes; the consequences of varying these parameters is discussed in more detail below.

The area graphs show the proportion of stimulus elements in A1 and A2 (left ordinate) during the first 200 timeslices of the trial for the absent target cue X (top row), the presented treatment cue T (middle row) and the outcome (bottom row). Recall that the cues were presented during Timeslices 1–10 and the outcome, if present, during Timeslices 11–20. These graphs represent only the initial part of the trial, which continued until all activity had decayed. The unshaded areas represent A1 activity, and the shaded areas (dotted lines where occluded) represent A2 activity. The net strength of the association between each cue node and the outcome node (right ordinate) during the trial is shown by the line graph. The associative strength is measured in terms of change from the strength for that cue at the end of compound TX training (therefore all associative strengths start at zero), with a change of 1.0 corresponding to a change equivalent to the mean increase in associative strength of cue X during the preceding six trials of TX^+ compound training using these parameters.

Unovershadowing. The process underlying unovershadowing can be clearly seen in the left-hand panel of Figure 1. The presentation of treatment cue T evoked the elements of both target cue X and the outcome into A2 through the associations formed during the prior TX^+ compound training. This conjoint A2 activity led to an increase in the associative strength of cue X across the trial with the size of the increments progressively declining as A2 activity decayed. The associative strength of cue T was initially reduced by simultaneous A1 activity in its elements and A2 activity in the outcome elements. Later in the trial, there was some recovery in the associative strength of cue T when the activity of its elements had decayed into A2, thereby yielding conjoint A2 activity with the outcome elements. However, the net outcome of the trial was a reduction in the associative strength of the treatment cue T, leading to extinction.

Backward inhibition. By contrast to unovershadowing, the associatively evoked A2 activity in cue X elements was initially paired with A1 activity in the outcome elements as soon as the outcome was presented on the first T^+ trial of the backward inhibition (TX^-T^+) contingency, as illustrated in the right-hand panel of Figure 1. The conjoint A2 A1 activity produced a large initial decrement in the associative strength of cue X. This decrement was counteracted later in the trial by an increment in associative strength generated by the conjoint A2 A2 activity when the majority of outcome elements had decayed from A1 to A2. However, this increment was not sufficient to reverse fully the initial decrement so that the target cue X had a net negative associative strength by the end of the trial. The treatment cue T showed increments in associative strength throughout the trials initially generated by the conjoint A1 A1 activity with the outcome elements and later by their conjoint A2 A2 activity.

Backward blocking. As is illustrated in the middle panel of Figure 1, essentially the same pattern of activity and associative change was seen under the backward blocking (TX^+T^+) contingency, except for the fact that the later increment in the associative strength of the tar-

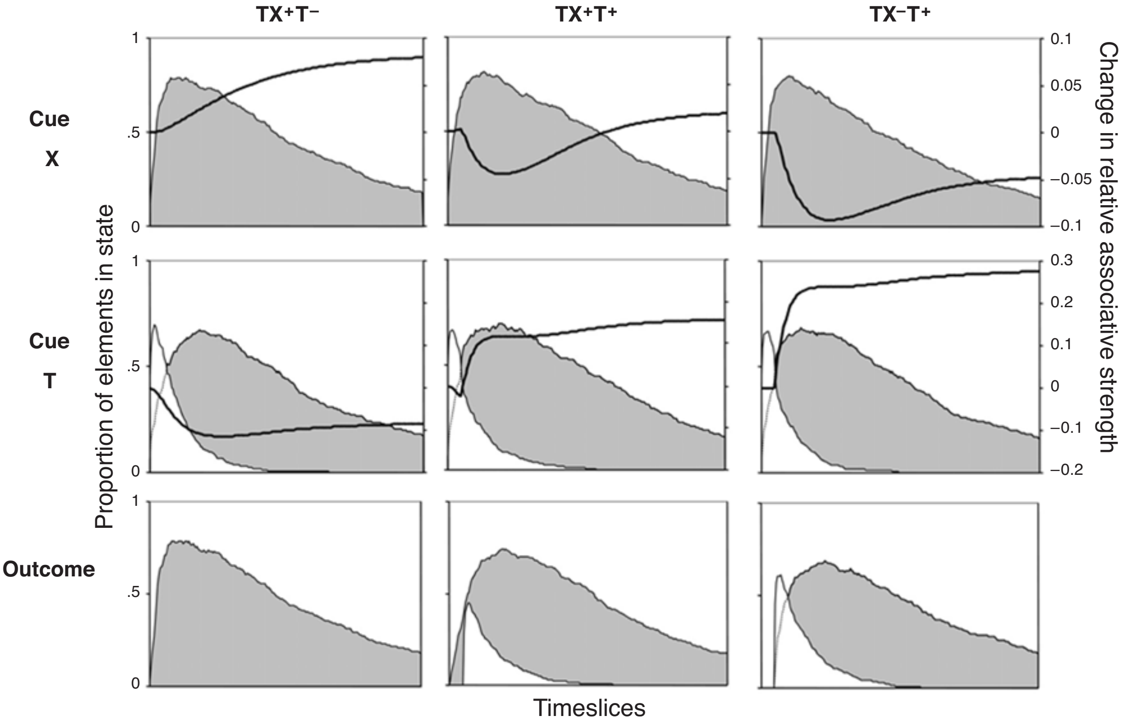

Figure 1. Simulation of the proportion of elements (left ordinate) in A1 (unshaded areas) and A2 (shaded areas) during the first 200 timeslices initiated by the first trial with the treatment cue T after six trials of TX compound training for the unovershadowing (TX^+T^-; left panel), the backward blocking (TX^+T^+; middle panel), and backward inhibition (TX^-T^+; right panel) contingencies. These proportions are shown separately for the target cue X (top row), the treatment cue T (middle row), and the outcome (bottom row). The treatment cue T was present during Timeslices 1–10 and the outcome, when presented, during Timeslices 11–20. The trials continued longer than shown, until all elements in all stimuli had decayed to inactive. The line graphs illustrate the changes in associative strength (right ordinate) of a cue from the strength acquired after the six trials of TX compound training. Note the different scale values for the target cue X and the treatment cue T. For further details, see the text.

get cue X almost completely canceled the initial decrement so that there was little net change in strength by the end of the trial. The difference was due to the fact that the treatment cue T had been pretrained as a predictor of the outcome in the backward blocking (TX^+T^+) contingency but not in the backward inhibition (TX^-T^+) contingency. As a consequence, presentation of cue T in the former case associatively retrieved some of the outcome elements into A2 prior to the presentation of the outcome, which enhanced the amount of conjoint excitatory A2 A2 activity while reducing the conjoint inhibitory A2 A1 activity.

In summary, these simulations of MSOP on the first trial with the treatment cue essentially confirmed the descriptive predictions derived by Dickinson and Burke (1996) and Larkin et al. (1998). This trial produced a net increment in the association strength of the target cue X under the unovershadowing (TX^+T^-) contingency, a net decrease under the backward inhibition (TX^-T^+) contingency, and little change under the backward blocking (TX^+T^+) contingency.

Terminal Associative Strengths

Although these simulations established the basic retrospective revaluation effect following a single trial with the treatment cue T alone, Larkin et al. (1998) in fact administered three such trials before requesting causal judgments. Consequently, the next simulations derived the associative strength of the target cue X after three treatment cue T trials. In addition to the three retrospective revaluation contingencies, we also simulated MSOP under simple overshadowing (TX^+), as the control contingency for assessing the source of retrospective revaluation, and under the analogous forward contingencies, forward (T^+TX^+) blocking, and forward (T^+TX^-) inhibition, using the training conditions of Larkin et al. Finally, we included a simple acquisition (X^+) contingency in which the treatment cue X alone was followed by the outcome for six trials.

In these simulations, all except one of the following parameters were set at the standard values used in the simulation of the first trial activity profiles: ρ ratio of A2 to A1 learning rates) = 0.1; l (the overall learning rate) = 0.01; N_d (the rate of decay of A1 activity) = 50 ; $P_{I\rightarrow A1}$ (salience) = .05; N_s (stimulus node size) = 400. A single critical parameter was varied from 10% to 200% of the standard value in 10% steps. Other parameters, namely the size of A1 and A2 states and the duration of the stimulus presentations, were not varied; such variation would be essentially equivalent to varying N_s (stimulus node size), or the ratio of $P_{\mathrm{I} \rightarrow \mathrm{A1}}$ (salience) to N_d (decay rate), respectively.

In order to compare the ordinal predictions of the model across different parameters, the final associative strengths of the other six contingencies were standardized as a proportion of the associative strength of target cue X from the overshadowing (TX$^+$) contingency. The simulations are presented as the average of three runs of the model under each parameter set.

Learning rates (ρ and l). Since we anticipated that the performance of MSOP would depend critically upon the ratio of the effectiveness of the different learning processes, we varied the ratio of learning rates when the elements of a cue were in A2 to the rate when its elements were in A1 (ρ) . Figure 2 shows the mean terminal relative associative strength for target cue X, relative to that from the overshadowing (TX$^+$) contingency, in each of the six other contingencies as a function of ρ.

For the simulations with a low ρ on the left side of Figure 2, the contribution of the learning processes engaged by A2 activity in cue elements was relatively small and the predictions of MSOP converged with those of standard SOP. With low ρ values, substantial forward learning effects were observed. Overshadowing was manifest by the fact that the relative associative strength for the acquisition (X$^+$) contingency was above +1, whereas the fact that the relative strength for the blocking (T$^+$TX$^+$) contingency was below +1 demonstrates the occurrence of simple forward blocking. Finally, under the inhibitory (T$^+$TX$^-$) contingency, the acquisition of inhibition is represented by the negative relative associative strengths acquired by target cue X.

These forward learning effects were, however, differentially sensitive to the increase in the second, A2-based learning processes resulting from an increase in ρ. Although there was a slight decline in the magnitude of overshadowing, the relative associative strength under the acquisition (X$^+$) contingency remained above +1 for all values of ρ. By contrast, the magnitudes of both forward blocking and inhibition progressively decreased when ρ was increased. These decrements reflected the enhanced excitatory learning between target cue X and the outcome, because of the conjoint A2 activity in both nodes toward the end of the compound trials.

Importantly, increasing ρ also introduced retrospective revaluation. Unovershadowing, which is represented by relative associative strengths for the unovershadowing (TX$^+$T$^-$) contingency above +1, emerged at low ρ values and remained sustained across the parameter range. Similarly, the relative associative strength for backward inhibition (TX$^-$T$^+$) dropped below zero as ρ was increased, although the magnitude of the backward inhibition effect was never very large under these training conditions. By contrast, backward blocking was never observed throughout the ρ parameter range using these schedules in that the relative associative strength for the backward blocking (TX$^+$T$^+$) contingency never dropped below +1.

In general, the predictions of MSOP in the range of ρ values between 0.05 and 0.15 were consistent with the ordinal pattern of results reported by Larkin et al. (1998).

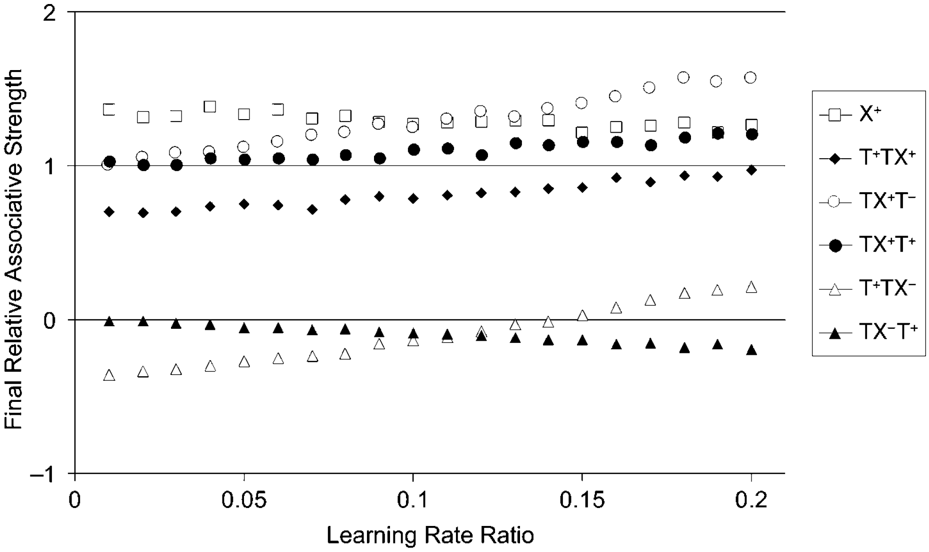

Figure 2. The simulated relative associative strength of the target Cue X after training on the various forward and backward contingencies as a function of the learning rate ratio (ρ). Training consisted of six TX compound trials and three treatment cue T trials, except for the acquisition (X$^+$) contingency, which consisted of six trials in which the target cue X was paired with the outcome. An associative strength of 1.0 is the strength acquired by the target cue X after six TX$^+$ compound trials. For further details, see text.

Although not presented graphically, it should be noted that increasing the overall learning rate parameter, l, enhanced the effects of both forward and backward contingencies without any substantial effects upon the ordinal predictions of the model. In the case of the backward blocking contingency (TX^+T^+), this effect took the form of augmenting the associative strength of the target cue X, presumably by enhancing the contribution of A2 A2 learning.

Decay rate (N_d). As illustrated in Figure 3, variation in the rate of decay (in terms of the number of null elements inserted during each timeslice) had a marked impact on the predictions of MSOP. At very low decay rates ($N_d \leq 10$), the predictions are clearly at variance with the observed effects. There was no retrospective revaluation in that backward blocking (TX^+T^+) and unovershadowing (TX^+T^-) yielded similar, high associative strengths, and there was no backward inhibition (TX^-T^+) and, indeed, no overshadowing.

Importantly, however, the observed pattern of forward and backward effects was robustly produced at intermediate decay rates. Backward inhibition was present at a similar magnitude for all decay rates above $N_d = 30$. Most importantly, retrospective revaluation was consistently predicted, with the final associative strength being larger for the target cue from the unovershadowing (TX^+T^-) contingency than that from the backward blocking (TX^+T^+) contingency. The highest decay rates did in fact yield associative strengths for the backward blocking (TX^+T^+) contingency that were numerically below +1, but the magnitude of this backward effect was very small. Moreover, at these decay rates the amount of forward blocking was small and the overall magnitude of retrospective revaluation was minimal due to a reduction in associative strength under the unovershadowing (TX^+T^-) contingency.

In summary, intermediate decay rates of around 50 null elements per timeslice yielded predictions that corresponded closely to the effects observed by Larkin et al. (1998).

Simulations under the final two parameters, stimulus node size (N_s) and stimulus salience ($P_{I \rightarrow A1}$), did not produce informative predictions and therefore are not depicted graphically. Increasing salience, not surprisingly, just enhanced the magnitude of both the forward and backward effects, whereas the predominant impact of increasing the stimulus node size was to reduce the magnitude of the forward blocking and inhibition effects. Variation of these parameters never yielded backward blocking.

Acquisition Functions

Although MSOP predicted the cardinal features of forward and backward learning, it is also important to establish that MSOP generates the orderly acquisition functions observed by Larkin et al. (1998). Consequently, we derived the relative associative strengths for both the treatment cue T and the target cue X during training on each of the contingencies described above. In order to demonstrate that the model is capable of predicting orderly extinction, as well as acquisition, the six acquisition trials from the X^+ contingency were followed by three extinction (X^-) trials.

Figure 4 shows a set of acquisition functions from single simulations of each contingency with the standard parameter set: $l = 0.01$; $\rho = 0.1$; $N_d = 50$; $P_{I \rightarrow A1} = .05$; $N_s = 400$. These functions are represented by the

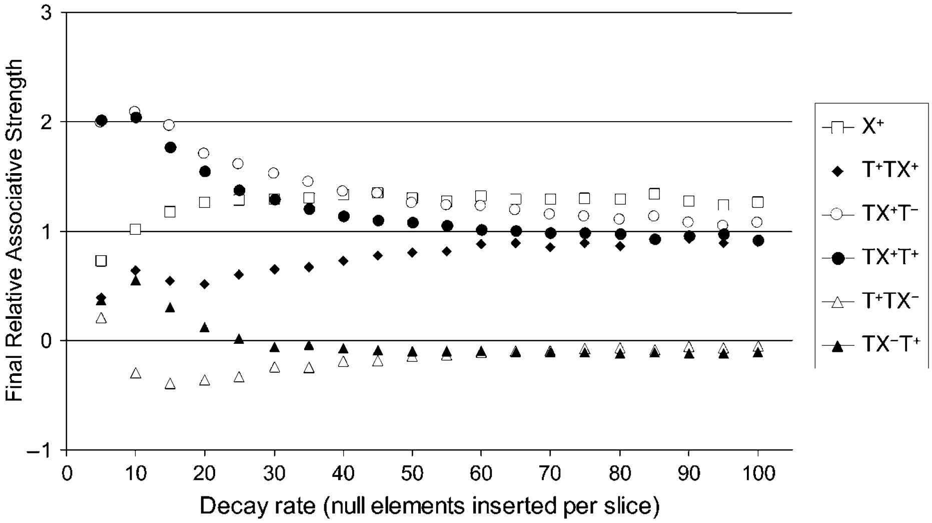

Figure 3. The simulated relative associative strength of the target cue X after training on the various forward and backward contingencies as a function of the decay rate at which null elements are inserted into A1 per timeslice. Training consisted of six TX compound trials and three treatment cue T trials, except for the acquisition (X^+) contingency, which consisted of six trials in which the target cue X was paired with the outcome. An associative strength of 1.0 is the strength acquired by the target cue X after six TX^+ compound trials. For further details, see text.

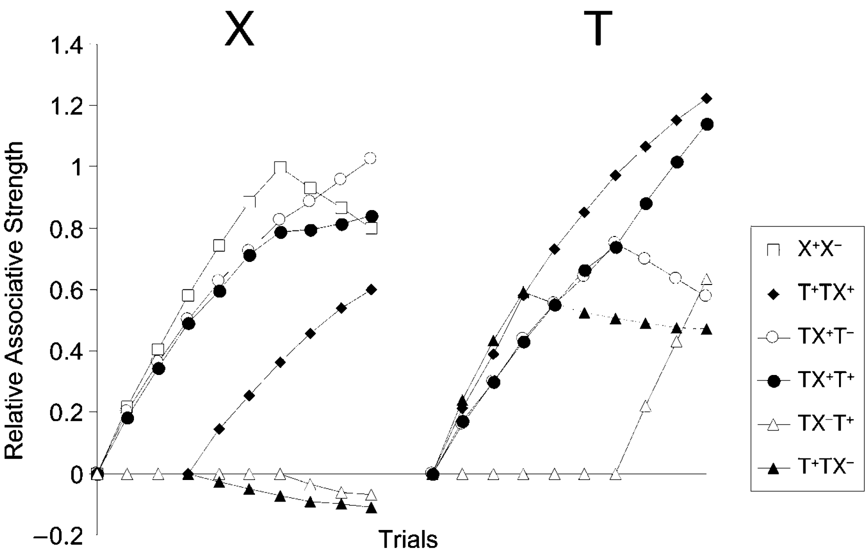

Figure 4. Relative associative strengths of the treatment cue T and the target cue X under various contingencies as a function of training trials. Training consisted of six TX compound trials and three treatment cue T trials, except for the acquisition extinction (X^+/X^-) contingency, which consisted of six trials in which the target cue X was paired with the outcome, followed by three trials in which the target cue X was presented alone. An associative strength of 1.0 is the strength acquired by the target cue X after six X^+ compound trials. For further details, see text.

relative associative strength of a cue where a strength of 1.0 is that acquired by the target cue X after the first six trials of the acquisition-extinction (X^+X^-) contingency.

Simply pairing a cue with an outcome produced a mildly negatively accelerated acquisition function, and a transition to the omission of the outcome for the last three trials of the acquisition-extinction (X^+X^-) contingency resulted in a loss of excitatory strength. Forward blocking is represented by the fact that the associative strength of the target cue X from the T^+TX^+ contingency on the final trial was less than that after six trials of TX^+ compound training under the retrospective TX^+T^+ and TX^+T^- contingencies. Unovershadowing was demonstrated by the increase in associative strength of target cue X during the final three trials of the TX^+T^- contingency. The fact that these trials, if anything, also produced an increase under the TX^+T^+ contingency illustrates the absence of a backward blocking effect.

DISCUSSION

Simulations of MSOP confirmed the qualitative predictions derived in the introduction. Both simple overshadowing and unovershadowing were predicted by the model across the parameter space. The acquisition (X^+) and the unovershadowing (TX^+T^-) contingency consistently yielded higher associative strengths for the target cue X relative to the overshadowing (TX^+) contingency. Indeed, learning under the acquisition and unovershadowing contingencies converged when the A2 learning rate was relatively high. Although the magnitude of the blocking effect in the forward blocking (T^+TX^+) contingency was inversely related to the relative learning rate, the effect was robust across variations in this parameter except at high values. By contrast, within the parameter set explored in this simulation, MSOP yielded no evidence for backward blocking under the TX^+T^+ schedule based on the training conditions of Larkin et al. (1998). The model also predicted inhibitory learning as manifest by net negative associative strengths across a range of parameters. The magnitude of forward inhibitory learning under the T^+TX^- contingency was inversely related to the relative learning rate and absent at high ratios. By contrast, backward inhibition under the TX^-T^+ contingency, although never large in magnitude, was positively related to the relative learning rate and absent at low values. There was, however, an intermediate range of learning rates that yielded both backward and forward inhibitory learning of approximately similar but small magnitude as well as the other contingency effects. Finally, it should be noted that acquisition functions derived from MSOP have an orderly and negatively accelerated form that is characteristic of human causal learning (e.g., Dickinson & Burke, 1996; Larkin et al., 1998).

MSOP is not the only associative theory that can explain the pattern of causal judgments observed under the forward and backward contingencies addressed in our simulation. Four in particular offer alternative accounts of retrospective revaluation, the phenomena most problematic for standard associative theories.

Van Hamme and Wasserman

Van Hamme and Wasserman (1994) suggested a liberalization of the Rescorla and Wagner (1972) learning (RW) rule that allows for learning about absent cues. The standard RW rule proposes that the change in associative strength is given by

$$\delta V = \alpha\beta\left(\lambda - \sum V\right), \qquad (4)$$

where $\sum V$ in the sum of associative strengths of all the cues present during the learning episodes and λ is the value of $\sum V$ required to predict the outcome. Lambda has a positive value on occasions when the outcome occurs, but is zero when the outcome does not occur. Alpha and beta are learning rate parameters that reflect the associability of the cue and the outcome, respectively. Because learning is driven by the prediction error $(\lambda - \sum V)$, which encodes the degree to which the outcome is unpredicted, the RW rule provides an elegant explanation of overshadowing and forward blocking as well as forward inhibition (see Rescorla & Wagner, 1972). However, absent cues have zero α values, rendering retrospective revaluation outside the scope of the rule because no learning can accrue to absent cues. In order to allow learning about absent cues, Van Hamme and Wasserman suggested that α is negative on episodes when that cue is absent.

Table 3 illustrates how the Van Hamme and Wasserman (1994) revision of the RW rule (VHW rule) explains retrospective revaluation for the backward blocking and unovershadowing contingencies. In the case of backward blocking, the value of α for the absent target cue X is negative on T^+ episodes in Stage 2, whereas the prediction error $(\lambda - \sum V)$ is positive. The associative strength necessary to predict the outcome is shared between Cues T and X during Stage 1, so that the associative strength of cue T is less than λ at least on the initial T^+ training episodes of Stage 2. This positive error term in combination with a negative α for the absent target cue X produces a decrement in its associative strength, thereby yielding backward blocking. A similar application produces unovershadowing except for the fact that in absence of an outcome in Stage 2, the prediction error is negative, and hence the increment in the associative strength of the target cue X is positive.

Table 3
Application of Van Hamme & Wasserman's (VHW) Revision of the Rescorla-Wagner (RW) Rule to Retrospective Revaluation

Stage 1	Stage 2	α	$\beta(\lambda - \Sigma V)$	dV
		Backward Blocking		
TX^+	T^+	positive	positive	positive
	\			
	x	negative	positive	negative
		Unovershadowing		
TX^+	T^-	positive	negative	negative
	\			
	x	negative	negative	positive

Note—T, treatment cue; X, target cue; x, representation of target cue X retrieved via a within-compound association (\) with the treatment cue T; $^+$, outcome; $^-$, no outcome.

In general, the MSOP model and the VHW rule yield similar predictions in the canonical contingencies except for one possible contrast noted by Larkin et al. (1998). According to the VHW rule, the absolute magnitude of the prediction error $(\lambda - \sum V)$ during the second stage of the backward blocking contingency should always be at least as large as that generated by the unovershadowing contingency. In the case of the backward blocking (TX^+T^+) contingency, the prediction error on the initial T^+ episode is $(\lambda - V_T)$, whereas the equivalent prediction error for the first T^- episode from the unovershadowing (TX^+T^-) contingency is $(0 - V_T)$. Because the total associative strength acquired during the TX^+ compound training will never be greater than λ and shared equally between cues T and X when of the same salience, the associative strength of the treatment cue T will never be greater than half λ. Consequently, the absolute prediction error for the backward blocking contingency is always as large or larger than that for the unovershadowing contingency. By contrast, as we have already noted, our simulation of MSOP failed to generate any backward blocking, whereas the same parameters produced substantial unovershadowing.

We noted in the introduction that the empirical evidence on backward blocking is mixed, although it can be reliably demonstrated under certain conditions (e.g., De Houwer et al. 2002; Le Pelley & McLaren, 2001; Shanks, 1985; Wasserman & Berglan, 1998). To the best of our knowledge, only three studies have directly compared these effects. Larkin et al. (1998) failed to detect reliable backward blocking and reported that the absolute magnitude of the unovershadowing was reliably greater than that of backward blocking, although it must noted that interpreting this difference involves a comparison between intervals at different points of the rating scale for causal judgments. The remaining studies (Le Pelley & McLaren, 2001; Wasserman & Berglan, 1998) all reported reliable backward blocking, although neither study directly compared the absolute magnitudes of the two effects. Since the stimuli playing the roles of cue T and cue X were counterbalanced in the Larkin et al. experiments, the relative magnitudes of backward blocking and unovershadowing cannot be attributed to differences in cue salience, α, and therefore, in order to explain this difference, the VHW rule would have to assume that the learning rate parameter, β, was sufficiently higher for the omission of the outcome than for its presentation to cancel any difference in the prediction error generated by the two contingencies.

Finally, we should note that MSOP provides a more complete account of retrospective revaluation than the VHW rule. Our implementation simulated the acquisition of associations between the treatment cue T and the target cue X. By contrast, the VHW rule leaves the process by which an absent cue acquires a negative alpha value unspecified.

APECS

Le Pelley and McLaren (2001) applied McLaren's (1993) APECS model to the backward blocking and unovershadowing contingencies. APECS is a complex multilayer connectionist network that uses configural representations and weight freezing to protect against interference. The complexity of APECS precludes a simple descriptive application of the model to the retrospective revaluation contingencies, but Le Pelley and McLaren (2001) demonstrated by simulation that the model predicts the retrospective revaluation patterns observed not only in their own experiments but also in those reported by Dickinson and Burke (1996) and Larkin et al. (1998) including backward inhibition.

Within the context of retrospective revaluation, Le Pelley and McLaren challenged MSOP with a new mixed ($TX^{+}T^{+/-}$) contingency, which mixed the backward blocking and unovershadowing contingencies. During the second stage of this contingency, the treatment cue T was paired with the outcome on half of the episodes but presented without the outcome on the remaining episodes. On the basis of a descriptive analysis, Le Pelley and McLaren argued that MSOP predicts that the causal judgments from the mixed ($TX^{+}T^{+/-}$) contingency should be more similar to those from the unovershadowing ($TX^{+}T^{-}$) contingency than to those from the backward blocking ($TX^{+}T^{+}$) contingency. In contrast to this prediction, they reported that the difference between the target cue X judgments for the unovershadowing ($TX^{+}T^{-}$) and mixed ($TX^{+}T^{+/-}$) contingencies was reliably greater than the difference between judgments from mixed ($TX^{+}T^{+/-}$) and backward blocking ($TX^{+}T^{+}$) contingencies.

In order to validate Le Pelley and McLaren's (2001) descriptive prediction for MSOP, we simulated the three contingencies using training schedules based upon their Experiment 4 (10 trials TX^{+}; 8 trials T^{+}/T^{-}). The mean final associative strength to cue X over three simulations of the backward blocking ($TX^{+}T^{+}$) and unovershadowing ($TX^{+}T^{-}$) contingencies was compared with the mean of six simulations of the mixed contingency (three with a T^{+} trial first during the second stage, and three with a T^{-} trial first). With the standard set of parameters, these simulations did not confirm Le Pelley and McLaren's (2001) intuition about the prediction of MSOP for their mixed ($TX^{+}T^{+/-}$) contingency. If anything, the simulated differences are in the same direction as they observed. The relative associative strength difference for target cue X between the unovershadowing ($TX^{+}T$) and the mixed ($TX^{+}T^{+/-}$) contingencies was 0.16, whereas that between mixed and backward blocking ($TX^{+}T^{+}$) contingencies was 0.13.

In summary, at present, there are appear to be no strong empirical grounds to choose between MSOP and APECS.

The Comparator Model

Retrospective revaluation in the form of unovershadowing was first reported in the animal conditioning literature by Kaufman and Bolles (1981) and, as an explanation of such effects, Miller and colleagues (e.g., Miller & Matzel, 1988) developed a comparator model. As in the case of the associative theories that we have already considered, the comparator model assumes that during the first stage of the unovershadowing ($TX^{+}T^{-}$) contingency, the treatment cue T and the target cue X both form within-compound associations between themselves in addition to associations with the outcome representation. However, the conditioned response to the target cue X, and by analogy the causal judgment for this cue, is determined not by the absolute strength of its associative connection with the outcome representation, but rather by this strength relative to the associative strength of the treatment cue.

Specifically, presentation of the target cue X not only activates the outcome representation directly but also activates it indirectly by exciting the treatment cue T representation, which in turn also activates the outcome representation. The magnitude of the response to cue X is then determined by the magnitude of the direct activation of the outcome representation relative to its indirect activation. Unovershadowing arises from the fact that presentation of the treatment cue T alone during the second stage of the unovershadowing ($TX^{+}T^{-}$) contingency extinguishes the association between treatment cue T and the outcome, with the consequence that the indirect activation of outcome representation produced by the target cue X is reduced. Correspondingly, backward blocking under the $TX^{+}T^{+}$ contingency occurs because the T^{+} episodes strengthen the cue T^{-} outcome association, thereby reducing the relative strength of the direct activation of outcome representation by the target cue X.

The empirical wedge that can be driven between the comparator theory and acquisition-based learning accounts, such as MSOP, concerns the role of within-compound associations. MSOP assumes that within-compound associations play a role only in retrospective revaluation contingencies, such as unovershadowing and backward blocking, but not in forward contingencies, such as simple blocking. As we described in the introduction, MSOP, like SOP itself, attributes forward blocking to a reduction in learning about the presented target cue X, resulting from the fact that the outcome is predicted by the pretrained treatment cue T. By contrast, comparator theory assumes a role for within-compound associations in both forward and backward contingencies because the only mechanism by which the treatment cue T can modulate responding to the target cue X is through their within-compound associations. The empirical evidence favors MSOP in this respect. Dickinson and his colleagues (Aitken et al., 2000; Dickinson & Burke, 1996; Larkin et al., 1998) consistently found that manipulations designed to reduce within-compound learning attenuated the impact of the predictive status of the treatment cue T on causal judgments for the target cue X in backward contingencies but not in forward ones.

One form of retrospective processing that favors comparator theory, at least in its extended form (Denniston, Savastano, & Miller, 2001), is higher order revaluation. Under a higher order contingency, the participant is initially trained with two compounds, TM$^+$ and MX$^+$, each paired with the outcome, during the first stage. Cue M can be regarded as a mediating cue between the treatment cue T and the target cue X. Second-order retrospective revaluation occurs when the causal ratings for the target cue X depend upon whether or not the treatment cue T is paired with the outcome during the second stage. Both De Houwer and Beckers (2002a) and Melchers, Lachnit, and Shanks (2004) have reported that the ratings of cue X are higher if the treatment cue T is paired with the outcome, which is the opposite of the first-order retrospective revaluation observed, in this case, for the mediating cue M.

These findings lie outside the scope of MSOP, which predicts that the treatment of cue T during the second stage does not impact on the associative status of target cue X. Even if the TM$^+$, MX$^+$ training resulted in an associative link between T and X being formed such that presentation of T would associatively activate X (recall that the current implementation of MSOP does not allow A2 activity of cue M to activate the elements of cue X via a T→M→X within-compound associative chain), there are two reasons why MSOP does not support higher order retrospective revaluation.

First, the initial compound training should, if anything, produce an inhibitory association between cues T and X. The presence of the mediating cue M in the TM compound should activate the elements of cue X into A2 through the within-compound association formed on MX trials. Consequently, A2 activity in cue X elements is paired with A1 activity in cue T elements on the TM trials, the condition for forming an inhibitory association between cues T and X. Second, even if the presentation of cue T did evoke activity in the elements of cue X during the second stage, the representational psychology that MSOP inherits from SOP requires that this activity is of the A2 form. Hence, any learning on second-order retrospective revaluation trials should bring about similar rather than opposed forms of learning to the equivalent first-order retrospective revaluation trials.

Whether or not extended comparator theory provides an adequate account of higher order retrospective revaluation has been disputed. Without going into details, we can state that this theory assumes that not only is the response to the target cue X modulated by the strength of the comparator cue, in this case cue M, through within-compound associations, but the effectiveness of the comparator cue is itself modulated by other cues with which it is associated, specifically the treatment cue T. In agreement with this analysis, Melchers et al. (2004) have recently reported that the magnitude of retrospective revaluation in a higher order contingency was correlated with the strength of the relevant within-compound associations. What is problematic for extended comparator theory, however, is that as in the first-order case, the equivalent correlations for the forward contingencies were not reliable.

Rehearsal Theory

The differential role of within-compound associations in the forward and backward contingencies led Melchers et al. (2004) to resurrect an account of retrospective revaluation originally proposed by Chapman (1991). The basic idea is that the presentation of the target cue T during the second stage of first-order retrospective contingencies retrieves a memory of the prior training trials with the TX compound and that the memory of these trials is then rehearsed. This rehearsal process is assumed to produce an effective contingency in which remembered compound TX trials are intermixed with the trials with the target cue T alone. This intermixing then allows the RW rule to generate retrospective revaluation effects by transforming backward contingencies into forward ones at the psychological level. A simple extension of this idea to the higher order contingencies predicts the observed forms of retrospective revaluation.

Like the VHW rule, however, the associative-rehearsal model of retrospective revaluation must await a formal specification the role of within-compound associations in the rehearsal process and its interface with learning before meriting a full evaluation against MSOP and APECS.

Other associative theories derived from studies of animal conditioning have been applied to human causal learning. Several demonstrations indicate that human participants often configure compound cues in causal and predictive learning tasks (e.g., Shanks, Charles, Darby, & Azini, 1998; Williams, Sagness, & McPhee, 1994), which has led to the application of Pearce's (1987, 1994) configural theories. Another class of associative theories focuses on the role of attentional and associability processes in learning, and in their original exposition of the associative account of causal learning, Dickinson et al. (1984) demonstrated that the Pearce-Hall attentional theory (Pearce & Hall, 1980) predicted causal judgments under different contingencies. More recently, Le Pelley and McLaren (2003) have applied Mackintosh's (1975) associability theory to human causal learning. However, none of these theories have, as yet, been revised and developed to provide an account of retrospective revaluation (but see Kruschke & Blair, 2001).

In conclusion, the present simulations have demonstrated that MSOP is a viable associative account of causal learning under contingencies that induce selective learning and retrospective revaluation. It must be acknowledged, however, that the empirical data do not exist as yet to allow a clear discrimination between MSOP and alternative associative theories, such as VHW (Van Hamme & Wasserman, 1994) and APECS (Le Pelley & McLaren, 2001).

REFERENCES

Aitken, M. R. F., Larkin, M. J. W., & Dickinson, A. (2000). Reexamination of the role of within-compound associations in the retrospective revaluation of causal judgements. *Quarterly Journal of Experimental Psychology*, **53B**, 59-81.

Chapman, G. B. (1991). Trial order affects cue interaction in contingency judgment. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **17**, 837-854.

De Houwer, J., & Beckers, T. (2002a). Higher-order retrospective revaluation in human causal learning. *Quarterly Journal of Experimental Psychology*, **55B**, 137-151.

De Houwer, J., & Beckers, T. (2002b). A review of recent developments in research and theories on human contingency learning. *Quarterly Journal of Experimental Psychology*, **55B**, 289-310.

De Houwer, J., Beckers, T., & Glautier, S. (2002). Outcome and cue properties modulate blocking. *Quarterly Journal of Experimental Psychology*, **55A**, 965-985.

Denniston, J. C., Savastano, H. I., & Miller, R. R. (2001). The extended comparator hypothesis: Learning by contiguity; responding by relative strength. In R. R. Mowrer & S. B. Klein (Eds), *Handbook of contemporary learning theory* (pp. 65-117). Mahwah, NJ: Erlbaum.

Dickinson, A., & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgements. *Quarterly Journal of Experimental Psychology*, **37B**, 397-416.

Dickinson, A., Shanks, D. R., & Evenden, J. L. (1984). Judgement 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, NS: Dalhousie University Press.

Kaufman, M. A., & Bolles, R. C. (1981). A nonassociative aspect of overshadowing. *Bulletin of the Psychonomic Society*, **18**, 318-320.

Kruschke, J. K., & Blair, N. J. (2001). Blocking and backward blocking involve learned inattention. *Psychological Bulletin & Review*, **7**, 636-645.

Larkin, M. J., Aitken, M. R. F., & Dickinson, A. (1998). Retrospective revaluation of causal judgments under positive and negative contingencies. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 1331-1352.

Le Pelley, M. E., & McLaren, I. P. L. (2001). Retrospective revaluation in humans: Learning or memory? *Quarterly Journal of Experimental Psychology*, **54B**, 311-352.

Le Pelley, M. E., & McLaren, I. P. L. (2003). Learned associability and associative changes in human causal learning. *Quarterly Journal of Experimental Psychology*, **56B**, 68-79.

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. In *Proceedings of the Fifteenth Annual Convention of the Cognitive Science Society* (717-722). Hillsdale, NJ: Erlbaum.

Melchers, K. G., Lachnit, H., & Shanks, D. R. (2004). Within-compound associations in retrospective revaluation and in direct learning: A challenge for comparator theory. *Quarterly Journal of Experimental Psychology*, **57B**, 25-53.

Miller, R. R., & Matzel, L. D. (1988). The comparator hypothesis: A response rule for the expression of associations. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 22, pp. 51-92). San Diego: Academic Press.

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

Pearce, J. M. (1987). A model of stimulus generalization for Pavlovian conditioning. *Psychological Review*, **101**, 587-607.

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

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, **87**, 532-552.

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.

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgments. *Quarterly Journal of Experimental Psychology*, **37B**, 1-21.

Shanks, D. R., Charles, D., Darby, R. J., & Azini, A. (1998). Configural processes in human associative learning. *Journal of Experimental Psychology: Learning, Memory & Cognition*, **24**, 1353-1378.

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning & Motivation*, **25**, 127-151.

Wagner, A. R. (1981). SOP: A model of automatic memory processing in animal behavior. In N. E. Spear & R. R. Miller (Eds.), *Information processing in animals: Memory mechanisms* (pp. 5-47). Hillsdale, NJ: Erlbaum.

Wasserman, E. A., & Berglan, L. R. (1998). Backward blocking and recovery from overshadowing in human causal judgment: The role of within-compound associations. *Quarterly Journal of Experimental Psychology*, **51B**, 121-138.

Williams, D. A., & Docking, G. L. (1995). Associative and normative accounts of negative transfer. *Quarterly Journal of Experimental Psychology*, **48A**, 976-998.

Williams, D. A., Sagness, K. E., & McPhee, J. E. (1994). Configural and elemental strategies in predictive learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 694-709.

NOTE

1. The divisions by *r* in Equation 1 reflects the ratio of excitatory and inhibitory learning proposed by Wagner (1981). This ratio produces no systematic change in associative strength to any constantly presented stimulus, as the ratio of elements in each state will correspond to the ratio of A1 and A2 sizes.