

Toward a Unified Model of Attention in Associative Learning

John K. Kruschke

Indiana University

Two connectionist models of attention in associative learning, previously used to model human category learning, are shown to have special cases that are essentially equivalent to N. J. Mackintosh's (1975, Psychological Review, 82, 276–298) classic model of attention in animal learning. The models unify formulas for associative weight change with formulas for attentional change, under a common goal of error reduction. Error-driven attentional shifting accelerates learning of new associations but also protects previously learned associations from retroactive interference. The models are fit to data from a recent experiment in human associative learning (J. K. Kruschke & N. J. Blair, 2000, Psychonomic Bulletin & Review, 7, 636-645), which shows that blocking of learning involves learned inattention. The approach also provides a novel and unifying theory of latent inhibition (the preexposure effect) in terms of blocking. The discussion summarizes how the approach accounts for a variety of other "irrational" phenomena in associative learning, including base rate effects, perseveration of attention through relevance shifts, overshadowing, and the extrapolation of rules near exceptions. Elsevier Science

The central role of attention in learning has been emphasized repeatedly in accounts of both human and animal learning. At least as early as Lawrence (1949, 1950), theorists of animal learning have argued that animals learn which cues should be attended to. One type of phenomenon addressed by such attentional learning theories is transfer of learning: When subsequent learning involves cues previously learned to be relevant, the speed of learning is faster than when the

This research was supported in part by NIMH FIRST Award 1-R29-MH51572. For comments on previous versions of this article, I thank Nathaniel Blair, Michael Erickson, Michael Fragassi, Mark Johansen, Peter Killeen, Nicholas Mackintosh, John Pearce, Roger Ratcliff, Teresa Treat, and Peter Wood. Parts of this research were presented at the Eighth Australasian Mathematical Psychology Conference, Perth Australia, 29 November 1997; the 31st Annual Conference of the Society for Mathematical Psychology, Vanderbilt University, Nashville TN, 8 August 1998; the Twentieth Annual Conference of the Cognitive Science Society, University of Wisconsin at Madison, 3 August 1998; and the Economic and Social Research Council Seminar on Knowledge, Concepts, and Categories, University College London, UK, 12 August 1999.

Address correspondence and reprint requests to John K. Kruschke, Department of Psychology, 1101 E. 10th St., Indiana University, Bloomington, IN 47405-7007. E-mail: kruschke@indiana.edu. URL: http:// www.indiana.edu/~kruschke/.



new learning involves cues previously learned to be irrelevant. Typically in these theories, the attention to a cue modulates the cue's influence on the animal's immediate response. Importantly, moreover, attention also modulates the cue's utilization in associative learning, and therefore the attention expresses the cue's associability.

The well-known Rescorla and Wagner (1972) model of associative learning formalized the idea (e.g., Kamin, 1969) that associations are learned between cues and *surprising* outcomes. The model acknowledged that different cues might be attended to differently, and therefore each cue was allowed a different learning rate to express the cue's individual associability. Crucially, however, Rescorla and Wagner (1972) provided no theory or formula describing how cue-specific learning rates should be adjusted by experience.

In a classic paper, Mackintosh (1975) proposed specific formulas expressing the idea that attention to cues that have been learned to be relevant should increase, but attention to cues that have been learned to be irrelevant should decrease. "In Mackintosh's model, however, surprise does not act via a comparable discrepancy [as in the Rescorla–Wagner model] and its role within [the formula for attention change] is not readily interpretable in terms of processing mechanisms.... [The formula for attention change] can probably receive a psychological interpretation within a different framework" (Dickinson, 1980, p. 153). One goal of the present article is to provide a framework wherein Mackintosh's (1975) formulas for attention learning and for association learning derive from the same motivation, gradient descent on error.

In the human learning literature, attention was for many years at the core of theories of concept learning. According to many theories of concept learning, people learn what stimulus dimension to attend to, and then people learn what the correspondence is from the features of that dimension to the concept label. Trabasso and Bower (1968), for example, in their book entitled Attention in Learning, provide a history and overview of models and empirical work. More recently, the idea of attention has played an important role in theories of category learning by humans. Building on work by Shepard, Hovland, and Jenkins (1961) and by Medin and Schaffer (1978), Nosofsky's (1986) generalized context model (GCM) suggested that people distribute attention to dimensions such that intercategory differences and intracategory similarities are maximized. The GCM had no mechanism by which those dimensional attention values were learned, however. Such a mechanism was provided by Kruschke's (1992) ALCOVE model, a connectionist implementation of the GCM in which dimensional attention strengths are gradually learned across trials by gradient descent on error. Kruschke (1996b) showed that a variant of ALCOVE can address transfer of learning across relevance shifts. Erickson and Kruschke (1998, see also Kruschke & Erickson, 1994) described an expanded connectionist architecture, called ATRIUM, that combines the exemplar representation in ALCOVE with rule-like representations, together with a mechanism that gradually learns to attend to one representation or the other, depending on the stimulus. Kruschke (1996a) introduced rapid attention shifts in a connectionist model called ADIT to account for perplexing base rate effects. These rapid attention shifts were used in an extension of the ALCOVE model, called

RASHNL, by Kruschke and Johansen (1999) to account for a wide spectrum of results in probabilistic category learning. A second goal of the present article is to show that these connectionist architectures of attention learning in humans (Erickson & Kruschke, 1998; Kruschke, 1996a; Kruschke & Johansen, 1999) have special cases that are essentially equivalent to the model of attention learning in animals proposed by Mackintosh (1975).

The equivalence between the connectionist models of human learning and Mackintosh's model of animal learning comes at the cost of abandoning Mackintosh's explanation of latent inhibition, also known as the preexposure effect, which is a phenomenon that occurs when an animal is preexposed to a cue with no novel outcome, but subsequently the cue is paired with a consequence. Learning to associate the cue with the consequence is retarded, compared with animals that were not preexposed to the cue (Lubow & Moore, 1959). Whereas several theories, like Mackintosh's, have explained the preexposure effect in terms of learned attention (e.g., Lubow, 1989; Pearce & Hall, 1980; Schmajuk, Lam, & Gray, 1996), none of these theories has formulas for attention change motivated in the same way as formulas for association change. By way of contrast, a third goal of the present article is to explain the preexposure effect as error reduction, just as associative learning is error reduction. In this approach, the preexposure effect is treated essentially as a special case of blocking of associative learning (Kamin, 1969). Recent data (Kruschke & Blair, 2000) are summarized which demonstrate that the associability of a redundant relevant cue is weakened in blocking, just as the associability of a preexposed cue is weakened in the preexposure effect. The data are well fit by the models. It must be stated at the outset, however, that this treatment of the preexposure effect is preliminary and motivated at this point primarily by theoretical symmetry. The main emphasis of this article is the connectionist models of human attentional learning and their relation to Mackintosh's (1975) model; the treatment of the preexposure effect is put forward tentatively.

Aside from the three aforementioned goals of this article, another theme is that shifts of attention during learning accomplish two complimentary effects: New learning is accelerated and previous learning is protected. These dual benefits are accomplished because attentional shifting reduces interference between old and new learning, and this reduction of interference is a natural consequence of error reduction.

Outline of Article

The next section describes a new connectionist model of associative learning, called EXIT, which combines rapidly shifting attention (Kruschke, 1996a; Kruschke & Johansen, 1999) with exemplar-specific attentional learning. The model is then fit to recent data which show that blocking of associative learning involves learned attention (Kruschke & Blair, 2000).

A second connectionist model is then described, based on a *mixture of experts* architecture (Erickson & Kruschke, 1998; Jacobs, Jordan, Nowlan, & Hinton, 1991; Kruschke & Erickson, 1994). This model also combines rapidly shifting attention with exemplar-specific attentional learning, but treats each cue as a separate

"expert" that individually attempts to predict the outcomes, without summing predictions from the other cues. This model is also fit to the blocking data.

Special cases of the two models are then shown to correspond very closely to the formulas presented by Mackintosh (1975) to account for animal learning. Unfortunately, the connectionist models cannot address the preexposure effect in the same way that Mackintosh (1975) suggested, so a new interpretation of the preexposure effect is suggested within the framework of the connectionist models.

The final discussion summarizes the ability of these types of models to address a number of other phenomena based on attentional shifts and learning. These phenomena highlight various consequences of attentional shifts: accelerated learning when few dimensions are relevant, perseveration of learned attention, protection of previous learning in base rate effects, and shifts of attention between rule-like and exemplar representations.

EXIT: THE EXTENDED ADIT MODEL

An essential aspect of the ADIT model (Kruschke, 1996a) is that each cue is multiplicatively gated by an individual attentional strength. A cue's attentional strength modulates (a) the cue's influence on immediate responding and (b) the cue's associability for imminent learning. On any given trial of learning, the attention strengths are shifted rapidly in response to error before the associative weights are adjusted. Attention is shifted away from cues that cause error and toward cues that reduce error. After this shift, the associative weights from the attended-to cues are adjusted (proportionally to the attention on each cue). Thus, the associative weights are affected only by cues that are attended to, and the model attends predominantly to those cues that reduce interference with previously learned knowledge. The attention shift thereby protects previously learned associations while accelerating the learning of new associations.

The original ADIT model shifted attention on each trial in response to error, but the shift was not retained in subsequent trials. Instead, attention was reset to default values at the beginning of each trial. This lack of learning about attention was not a theoretical commitment, but was only a convenience to reduce the number of free parameters, because attentional learning was not needed to address the empirical phenomena accompanying ADIT's original article (Kruschke, 1996a). On the contrary, the gradual learning of attention has always been an underlying commitment for the approach (e.g., Erickson & Kruschke, 1998; Kruschke, 1992, 1996b; Kruschke & Johansen, 1999), and this commitment will be honored in the present extension of ADIT. The extended version is called EXIT. The name "EXIT" is mnemonic in three ways. First, it is short for *Ex*tended AD*IT*. Second, it is short for *Ex*emplar-based attention to distinctive *input*. Third, an ADIT is an entrance, and an EXIT is encountered later.

Activation Propagation to the Category Nodes

Figure 1 shows the architecture of EXIT. Each component cue in the stimulus is represented by a corresponding input node in a connectionist network, and each

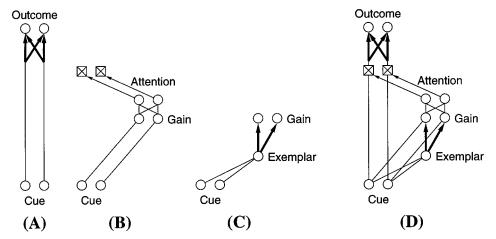


FIG. 1. Architecture of the EXIT model. This diagram illustrates a case with two cues, one exemplar, and two outcomes. (A) The basic connections from cues to outcomes are shown: Thick arrows denote learnable associative weights, denoted by w_{ki} in Eq. (1). (B) The network mechanism for allocating default, normalized attention to cues is shown. The activation of the gain nodes is expressed by Eq. (4). The criss-crossing lines from gain nodes to attention nodes represent the normalization of attention expressed by Eq. (5). The X's in boxes directly above the input cues represent the multiplicative application of the attention on the cues, as expressed in Eq. (1). (C) The network mechanism for learning new attentional distributions is shown. The activation of an exemplar node is specified in Eq. (3). The thick arrows from the exemplar to gain nodes represent learned associative weights, denoted w_{ix} in Eq. (4). (D) The complete architecture, with the components from panels (A), (B), and (C) superimposed, is shown.

possible outcome is represented by a corresponding output node. If cue i is present in a stimulus, then node i is activated, with activation value $a_i^{\rm in}=1$. When cue i is absent, $a_i^{\rm in}=0$. When outcome k is present, then the corresponding output node receives a "teacher" signal $t_k=1$, which indicates that the node should be activated. When the outcome is absent, then $t_k=0$. Input node i is connected to output node k via a link with an associative strength, or weight, denoted w_{ki} .

When a stimulus is presented, the corresponding input nodes are activated, and activation spreads to the output nodes via the weighted connections. The attentional strengths also modulate the influence of the input activations, such that the output activation is determined by a weighted sum across the attentionally gated input activations. Formally, the activation of the output node k is determined by

$$a_k^{\text{out}} = \sum_i w_{ki} \alpha_i a_i^{\text{in}}, \tag{1}$$

where α_i is the attention strength on the input node *i*. The source of these attention values will be described below. The input-to-category association weights are initialized at zero, but change with learning, as described later. The path of activation from cues through associative weights to outcomes is shown in Panel (A) of Fig. 1.

Category node activations are mapped to response probabilities using a version of the Luce (1959) choice rule, also known as the "softmax" rule (Bridle, 1990;

Rumelhart, Durbin, Golden, & Chauvin, 1995). Specifically, the probability of choosing category c is given by

$$p(c) = \exp(\phi a_c^{\text{out}}) / \sum_k \exp(\phi a_k^{\text{out}}), \tag{2}$$

where ϕ is a scaling constant. In other words, the probability of classifying the given stimulus into category c is determined by the magnitude of category c's activation relative to the sum of all category activations. The constant, ϕ , determines the decisiveness of the network: A large value of ϕ expresses a highly decisive choice, in that it causes just a small activation advantage for category c to be translated into a large choice preference for category c. A small value of ϕ expresses an indecisive or unconfident network, in that the small ϕ causes large activation differences to be translated into ambivalent choices.

This rule for mapping output activations to choice probabilities has many precedents in the psychological literature (e.g. Estes, 1988, 1994; Gluck & Bower, 1988a; Kruschke, 1992) and in the engineering literature (e.g., Bridle, 1990; Rumelhart *et al.*, 1995). An added computational benefit beyond the psychological plausibility is that exponentiation of the output activations monotonically transforms possibly negative activations into positive values, which is essential if the transformed values are interpreted as probabilities. Appendix 1 discusses an alternative mapping from activations to choice probabilities, in which the activations are raised to a power instead of exponentiated.

Base Rates. The original ADIT model (Kruschke, 1996a, p. 15, Eq. (11)) used a separate formula for mixing category base rates with the choice probabilities generated from the associative network. Appendix 2 shows that this separate formula is essentially equivalent to handling base rates as learned associations from a bias cue. The bias cue is fully activated on every trial. In effect, the bias cue encodes the response prompt that appears in every trial during an experiment. It is possible, however, for the bias cue to have a different salience than the other cues (cf. Kruschke & Johansen, 1999). It is also possible that the attention should not be as mutable on the bias cue as on the other cues.

Appendix 2 discusses some ramifications of learning with a bias cue. In any case, what was presented in the original ADIT model as a separate principle for mixing base rates with other choice probabilities is actually equivalent to the singular attentional and associative learning system applied to a bias cue. In the experiment fitted later in the article, the base rates of the categories were all equal (particularly in the final phase of training), and so the bias cue is omitted from the model for simplicity.

Activation Propagation in the Attentional System

Panels (B) and (C) of Fig. 1 illustrate the attentional system. This attentional system maps input activations to attention strengths. An assumption of the model is that total attentional capacity is limited, so that increasing attention to one cue entails reducing attention to other cues. This capacity limit is implemented in the

model by assuming that each cue has an underlying attentional *gain*, and these gains are normalized (in a way to be specified below) to generate attention strengths. The normalization is indicated in Panel (B) of Fig. 1 by the criss-crossing connections from gain nodes to attention nodes.

The architecture of the attentional system is designed to accomplish two distinct goals. The first goal is to implement the default assumption that any presented cue should get some attention. This goal is achieved by providing each gain node with a "hard-wired" connection from each corresponding cue. These hard-wired one-to-one connections are shown in Panel (B) of Fig. 1 as solid lines from the cues to the gain nodes.

The second goal of the attentional module is to learn how attention should be distributed over the stimulus cues as a function of the particular combination of cues. In principle, the mapping from stimulus cues to attention values could be highly nonlinear, and so the module should be given adequate computational capacity to learn nonlinear mappings. Perhaps the most straightforward architecture for accommodating nonlinear mappings involves exemplar mediation of the mapping from input to output. Therefore, the model recruits exemplar nodes whenever a novel stimulus configuration is encountered, and the connection weights from the exemplar nodes to the gain nodes learn to predict the shifted attentional gains. These adaptive weights are shown as thick arrows in Panel (C) of Fig. 1.

Exemplar-mediated mappings are reasonably motivated psychologically, as well as computationally. In everyday life, if we learn to ignore a cue in one situation, we don't necessarily ignore it in all situations. In experiments on human category learning, it has been shown that dimensions ignored for some stimuli are not ignored for other stimuli (e.g., Aha & Goldstone, 1990). Because some input cues can be context cues, the exemplar nodes in the attentional module can also encode contextual cues and thereby implement context specificity.

The activation of an exemplar node corresponds to the psychological *similarity* of the current stimulus to the exemplar represented by the node. Similarity drops off exponentially with distance in psychological space, as suggested by Shepard (1987), and distance is computed using a city-block metric for psychologically separable dimensions (Garner, 1974; Shepard, 1964). Each exemplar node is significantly activated by only a relatively localized region of input space; i.e., it has a small "receptive field." Formally, the activation value of exemplar x is given by

$$a_x^{\text{ex}} = \exp\left(-c\sum_i |\psi_{xi} - a_i^{\text{in}}|\right),\tag{3}$$

where the superscript "ex" indicates that this is an exemplar node; c is a constant called the *specificity* that determines the overall narrowness of the receptive field; and ψ_{xi} represents the presence or absence of cue i in exemplar x, such that $\psi_{xi} = 1$ if cue i is present in the exemplar and $\psi_{xi} = 0$ if cue i is absent. This is the same

¹ Alternatively, the exemplar nodes could form a random covering map of the input space, as in the original ALCOVE model (Kruschke, 1992). Perhaps the best option would be recruit exemplars in response to error. These options are not explored here.

exemplar-similarity function used in the ALCOVE model (Kruschke, 1992) and in the generalized context model (Nosofsky, 1986).

Within the attention module activation propagates from the input nodes to the gain nodes via two paths: along the previously described one-to-one connections from input nodes to gain nodes shown in Panel (B) of Fig. 1 and via exemplar nodes to gain nodes, shown in Panel (C) of Fig. 1. The activation of gain node i is given by

$$g_i = a_i^{\text{in}} \exp\left(\sum_x w_{ix} a_x^{\text{ex}}\right),\tag{4}$$

where w_{ix} is the associative weight from exemplar node x to gain node i. The weights in Eq. (4) are initialized at zero but change to new values with learning, as described below. Equation (4) gives zero gain to input cues with zero activation, and a gain of 1 to input cues about which nothing has been learned yet. Note also that the gains on all cues are nonnegative.

From the gain nodes, activation propagates to the attention nodes. The capacity constraint is formalized by requiring the length of the attention vector to be equal to 1, with length measured by a Minkowski power metric. Formally, this is denoted as the constraint that $\sum_i \alpha_i^P = 1$, where P > 0 is the value of the power in the Minkowski metric. Then the attention to the *i*th cue is just the normalized gain of the *i*th cue,

$$\alpha_i = g_i / \left(\sum_j g_j^P\right)^{1/P}.$$
 (5)

The denominator is certain to be greater than zero because the gains computed from Eq. (4) are nonnegative, and at least one gain is nonzero by design. Increased attentional capacity is reflected by larger values of P. When P=1, the attention strengths must sum to unity, and the attention to any one cue is just the proportion of its gain relative to the total of the other gains, i.e., $\alpha_i = g_i/\sum_j g_j$. In this case, any increase of attention to one cue comes at the cost of the same amount of decrease in attention to other cues. When the capacity P approaches infinity, the attention to each cue approaches the proportion of its gain relative to the maximal gain of any cue, i.e., $\alpha_i = g_i/\max_j \{g_j\}$. The cue with maximal gain gets an attentional strength of nearly 1, and other cues get attention proportional to the maximal gain. If several cues are tied for maximal gain, they all get attention of nearly 1. When 0 < P < 1, any increase in attention to a cue causes more than that amount of decrease to other cues; in this case there is severe competition for attention among cues, and there is relatively little attention to any cue unless all cues but one have attention strengths close to zero.

Attention Shifting

After activation is propagated to the category nodes and the categorization probabilities are determined, corrective feedback is supplied, just as in human

learning experiments. The first response to this corrective feedback is a rapid shift of attention to reduce error. Error is measured as the sum squared deviation between the teacher values and the generated activation values, across the output nodes; i.e.,

$$E = .5 \sum_{k} (t_k - a_k^{\text{out}})^2.$$
 (6)

The coefficient .5 appears in Eq. (6) only for convenience in subsequent derivations. This definition of error is typical for models that learn by gradient descent on error (Gluck & Bower, 1988b; Kruschke, 1992; Rumelhart, Hinton, & Williams, 1986), but other definitions of error are possible (Rumelhart *et al.*, 1995).

Attention is adjusted by gradient descent on error with respect to the underlying gains. As a preliminary step in deriving the gradient, the derivative of attention with respect to gain will be computed now. In this and all subsequent formulas, a lower case subscript denotes an index that can vary, whereas an upper case subscript denotes an index that has a fixed value. From Eq. (5), we find that the derivative of attention to some cue i with respect to the gain of specific cue I is

$$\frac{\partial \alpha_{i}}{\partial g_{I}} = \left[\left(\sum_{j} g_{j}^{P} \right)^{1/P} \kappa_{iI} - g_{i} \frac{1}{P} \left(\sum_{j} g_{j}^{P} \right)^{(1/P)-1} \left(\sum_{j} P g_{j}^{P-1} \kappa_{jI} \right) \right] / \left[\sum_{j} g_{j}^{P} \right]^{2/P} \\
= \left(\kappa_{iI} - \alpha_{i} \alpha_{I}^{P-1} \right) / \left(\sum_{j} g_{j}^{P} \right)^{1/P}, \tag{7}$$

where $\kappa_{iI} = 1$ if i = I and $\kappa_{iI} = 0$ otherwise, which is sometimes referred to as the Kronecker delta function of i and I. (The traditional notation for the Kronecker delta function, δ_{iI} , is avoided to prevent possible confusion with the delta rule in connectionist learning.) Then, applying the chain rule to Eq. (6), we find that gradient descent on error with respect to gains yields

$$\begin{split} \Delta g_{I} &= -\lambda_{g} \frac{\partial E}{\partial g_{I}} \\ &= \lambda_{g} \sum_{k} (t_{k} - a_{k}^{\text{out}}) \sum_{i} w_{ki} a_{i}^{\text{in}} \frac{\partial \alpha_{i}}{\partial g_{I}} \\ &= \lambda_{g} \sum_{k} (t_{k} - a_{k}^{\text{out}}) \sum_{i} w_{ki} a_{i}^{\text{in}} (\kappa_{iI} - \alpha_{i} \alpha_{I}^{P-1}) \Big/ \Big(\sum_{j} g_{j}^{P} \Big)^{1/P} \\ &= \lambda_{g} \sum_{k} (t_{k} - a_{k}^{\text{out}}) (w_{kI} a_{I}^{\text{in}} - \alpha_{I}^{P-1} a_{k}^{\text{out}}) \Big/ \Big(\sum_{i} g_{j}^{P} \Big)^{1/P}, \end{split} \tag{8}$$

where λ_g is a positive constant of proportionality called the *shift rate* for attention. Psychologically, attention is hypothesized to shift a large extent on a single trial. This large shift cannot be achieved formally with a single large step in the direction of the gradient because attention is a highly nonlinear function of gain; that is, the gradient changes as the attention changes. Therefore, the change specified by the equation for gain change is iterated 10 times (an arbitrary number) on each trial,

so that the nonlinearity of the function can be approximated with 10 relatively small steps. After each small attention change the activation is repropagated to the category nodes to generate a new error, and attention is changed a small amount again, for 10 iterations. (On any one of these iterations, if a gain value is driven to a negative value, it is simply reset to 0 before the attention values are computed.) The result of these 10 small steps constitutes the single large shift. The same method was applied in the RASHNL model (Kruschke & Johansen, 1999).

Learning of Associations

After the attention is shifted the association weights are adjusted, also by gradient descent on error,

$$\begin{split} \varDelta w_{\mathit{K}\!\mathit{I}} &= -\lambda_{\scriptscriptstyle{W}} \frac{\partial E}{\partial w_{\mathit{K}\!\mathit{I}}} \\ &= \lambda_{\scriptscriptstyle{W}} (t_{\scriptscriptstyle{K}} - a_{\scriptscriptstyle{K}}^{\mathrm{out}}) \, \alpha_{\mathit{I}} a_{\mathit{I}}^{\mathrm{in}} \end{split} \tag{9}$$

where λ_w is a constant of proportionality called the *learning rate for output weights*. The associative weights for the gain nodes are also adjusted via gradient descent on error, where error is defined as the sum of squared differences between the shifted value and the initial preshift value. That is, the shifted values act as the teachers for the gain node activations. Formally, this yields

$$\Delta w_{IX}^{g} = \lambda_{x} (g_{I}^{\text{shift}} - g_{I}^{\text{init}}) g_{I}^{\text{init}} a_{X}^{\text{ex}}, \tag{10}$$

where λ_x is the learning rate for the associative weights from the exemplar to gain nodes. For infinitesimal shifts, this change is equivalent to gradient descent on the output error, E.

List of Free Parameters in EXIT. The free parameters of the EXIT model are the following:

- 1. the response probability scaling constant ϕ , used for converting output activation to response probability, in Eq. (2);
 - 2. the specificity c of the exemplar nodes in the attention module, in Eq. (3).
- 3. the attention normalization power P, i.e., the attentional capacity, in Eq. (5),
 - 4. the attention shift rate λ_g in Eq. (8);
- 5. the associative weight learning rate λ_w for categorization module, in Eq. (9); and
- 6. the learning rate λ_x for the associative weights from exemplar nodes to gain nodes, in Eq. (10).

EXPERIMENT: BLOCKING INVOLVES LEARNED INATTENTION

One demonstration of the importance of attention in EXIT is its ability to account for attenuated learning about a previously blocked cue. Blocking of

associative learning was first reported by Kamin (1968) and can be described as follows: Consider two cues, A and B, that are always followed by outcome 1. This correspondence is denoted $AB \rightarrow 1$. On average, the two cues each gain some positive associative strength with the outcome. In contrast, when $AB \rightarrow 1$ is preceded by an earlier phase of training without B (i.e., $A \rightarrow 1$), then the subsequent training with A and B together seems to generate little learning about B. The previous training with A alone has apparently prevented, or *blocked*, subsequent learning about the redundant relevant cue, B.

Blocking is a historically crucial finding because it disconfirms all models of learning in which associative strength is incremented by the mere contiguity of cue and outcome. This is because Cue B and outcome 1 co-occurred many times yet there was apparently little associative strength built up. "No empirical finding in the study of animal learning has been of greater theoretical importance than the phenomenon of blocking" (Williams, 1999, p. 618). The dominant explanation of blocking was suggested by Kamin (1968) and formalized in the classic model of Rescorla and Wagner (1972). The idea is that associative strength changes only to the extent that the outcome is unexpected. The Rescorla-Wagner model is essentially a special case of EXIT when the attentional system is excised and what remains is the cue-to-outcome associations. Output activation is defined as in Eq. (1) and weight changes are defined as in Eq. (9). The discrepancy between teacher value and generated value, $t_k - a_k^{\text{out}}$ in Eq. (9), expresses the degree to which the outcome is unexpected. The Rescorla-Wagner model implies that little is learned about the redundant relevant cue because the occurrence of the outcome is fully predicted by the first cue. The Rescorla-Wagner model acknowledged that different input cues could have different associabilities, but provided no mechanism whereby these associabilities change due to training. The Rescorla-Wagner model has had a monumental influence on research in associative learning (Miller, Barnet, & Grahame, 1995; Siegel & Allan, 1996), and it remains the standard explanation of blocking (e.g., Domjan, 1998, pp. 107–110).

An alternative explanation of blocking was propounded by Mackintosh and colleagues (Mackintosh, 1975; Mackintosh & Turner, 1971; Sutherland & Mackintosh, 1971). Their *attentional* theory suggested that something *is* learned about the redundant relevant cue, namely that it should be ignored. Mackintosh (1975) proposed specific formulas to govern changes in attention, and it will be shown later in this article that a special case of EXIT yields attentional changes essentially the same as those proposed by Mackintosh (1975).

This attentional explanation of blocking also implies that a blocked cue should subsequently suffer attenuated learning, because the suppression of attention must be unlearned. Mackintosh and Turner (1971) reported just such attenuation after blocking in rats. Recently, experiments with human participants also showed that learning about a blocked cue is attenuated relative to a control cue (Kruschke & Blair, 2000). One of these experiments (Kruschke & Blair, 2000, Experiment 1) is summarized here, and then EXIT is fit to the data. The fit is reasonably good, but the fit is significantly worse and qualitatively lacking when the attentional shifts are fixed at zero and the model only implements error-driven learning of associative weights. That is, attentional shifts are needed for a full account of blocking.

Design of the Experiment

Participants learned to diagnose lists of symptoms as various diseases. On a given learning trial, the participant would see a list of symptoms (e.g., "ear ache" and "skin rash") on his or her computer screen and have to diagnose this hypothetical patient as having one of six diseases, D, F, G, H, J, or K, by pressing the corresponding key on the computer keyboard. After making his or her response, the participant saw corrective feedback. Initially, the participant would merely guess, but after several trials would learn the correct diagnoses. See Kruschke and Blair (2000) for complete procedural details.

The experiment consisted of three phases of training, shown in Table 1. In the first phase, symptom A always resulted in disease 1, denoted $A \rightarrow 1$. (Please note that the abstract specification uses letters to denote symptoms and numerals to denote diseases, unlike the actual stimuli presented to participants.) In the second phase of training, the redundant symptom B was added to symptom A, always leading to the same disease as that which previously occurred with symptom A (i.e., $AB \rightarrow 1$), so that learning about symptom B would, presumably, be blocked. The second phase also included $HI \rightarrow 6$, which acted as a comparison for the blocked symptom B. In the test phase for blocking, symptoms were presented without corrective feedback. Of several cases tested, one was a combination of symptoms B and I. If symptom B was blocked, then people should prefer the disease paired with the control symptom I over the disease paired with the blocked symptom B.

In the subsequent, third phase of training, new symptoms and diseases were introduced, such that $ABC \rightarrow 2$ and $DEF \rightarrow 4$. The central motivations for this structure are the hypotheses that (1) learners will shift attention away from cues that already have been learned as indicative of different diseases (Kruschke, 1996a; Kruschke & Johansen, 1999) and (2) learners will tend not to shift attention toward a cue that they have previously learned to ignore. For the case $DEF \rightarrow 4$, attention will shift away from symptom D, because it is already known to indicate disease 3, leaving attention on the distinctive symptoms E and F. For the case $ABC \rightarrow 2$,

TABLE 1

Design of Experiment Showing Attenuation after Blocking (Kruschke & Blair, 2000, Experiment 1)

Phase	Blocking	Control to assess attenuation	Control to assess blocking
Training I	$A \rightarrow 1$	$D \rightarrow 3$	
Training II	$AB \rightarrow 1$	$D \rightarrow 3$	$HI \rightarrow 6$
Test for blocking		e.g., BH, BI	
Training III	$A \rightarrow 1$	$D \rightarrow 3$	$G \rightarrow 5$
	$ABC \rightarrow 2$	$DEF \rightarrow 4$	$GHI \rightarrow 6$
Test for attenuation		e.g., BE, BF	

Note. Letters A-I denote symptoms, and numerals 1-6 denote diseases.

attention will shift away from symptom A, because it is already known to indicate disease 1. If, as a consequence of blocking, people have learned to ignore symptom B, then attention should also not be directed to Symptom B, leaving only symptom C to be significantly attended to. Then there will be only a relatively weak association made between symptom B and Disease 2. The strength of the association is assessed in the final test phase, when symptoms B and E are presented together. It was predicted for this case that people would prefer the disease paired with the nonblocked symptom better than the disease paired with the blocked symptom.

On the other hand, if during the second phase, when symptom B is blocked, there is no learned inattention to symptom B, and instead there is merely a relative lack of associative learning about symptom B, then subsequent learning about it should be largely unaffected. That is, according to the Rescorla–Wagner (1972) model, associative learning from B to the novel disease 2 should be unaffected by any previous (lack of) learning from B to disease 1. Therefore, in the third training phase, symptom B should be as strongly associated with the new disease 2 as the control symptom E is associated with the new disease 4. A variety of other symptom combinations was presented in the final testing phase to further constrain the theories.

Table 1 shows that the third training phase also included cases of symptom G paired with outcome 5 and symptoms GHI paired with outcome 6. These cases were included merely to match the third phase of training in the other experiments reported by Kruschke and Blair (2000), in order to facilitate comparison of results across experiments.

Results of the Experiment

Table 2 shows the choice proportions for each of the test cases. The most critical elements of the results are set in boldface font in the table and graphed in Fig. 2. Robust blocking is exhibited for the test case BH/BI: People preferred disease 6, associated with the control symptoms H and I, over disease 1, associated with the blocked symptom B, 58.8% to 15.0%, χ^2 (df = 1, N = 59)/2 = 10.4, p < .005. The χ^2 value has been divided by 2 as the most conservative precaution against a possible lack of independence between the two repetitions of the case seen by each participant (Wickens, 1989, p. 28).

An interesting aspect of the data from the first test phase (the test for blocking) is that people often selected diseases they had not yet seen any cases of. For example, the test case AD elicits a total of 27.6% responses for Diseases 2, 4, or 5, none of which had yet occurred in training. A possible explanation is that people were using what I have previously referred to (Kruschke & Bradley, 1995; Kruschke & Erickson, 1995) as *strategic guessing*, whereby people might reason that "this is a case I haven't seen before, therefore it must be a disease I haven't seen before." The models have no mechanism for strategic guessing. This strategic guessing cannot provide an alternative explanation of the effect attributed to blocking, however, because the preferred responses involved the already learned diseases.

TABLE 2

Human Choice Percentage in Test Trials of Experiment 1 of Kruschke and Blair (2000)

			Dis	ease		
Symptoms	1	2	3	4	5	6
		Test	for Blocking			
вн/ві	15.0	6.3	3.8	6.3	10.0	58.8
AB	81.3	3.8	3.8	3.8	5.0	2.5
D	2.5	0.0	96.3	1.3	0.0	0.0
HI	1.3	0.0	0.0	1.3	2.5	95.0
BD	15.0	3.8	66.3	6.3	2.5	6.3
AD	42.5	13.8	30.0	2.5	11.3	0.0
AH/AI	65.0	1.3	1.3	5.0	3.8	23.8
DH/DI	2.5	6.3	43.8	5.0	10.0	32.5
		Test fo	or Attenuation			
BE/BF	2.5	22.5	2.5	58.1	3.1	11.3
A	94.4	1.3	1.3	1.3	0.6	1.3
ABC	13.8	72.5	0.6	4.7	2.5	5.0
D	3.1	0.6	93.8	1.3	0.0	1.3
DEF	1.3	5.0	12.5	72.5	2.5	6.3
G	0.6	0.0	1.9	0.6	95.6	1.3
GHI	1.3	6.3	2.5	5.6	6.9	77.5
CE/CF	1.3	39.4	5.0	42.5	3.1	8.8
BH/BI	1.3	21.9	0.6	6.3	3.8	66.3
CH/CI	1.9	48.8	1.9	3.1	5.0	39.4
AB	56.9	27.5	3.1	4.4	0.6	7.5
AC	34.4	56.3	1.3	1.3	3.1	3.8
DE/DF	1.3	6.3	31.3	51.9	3.1	6.3
GH/GI	0.6	4.4	3.8	3.8	34.4	53.1

Note. Letters A–I denote symptoms, and numerals 1–6 denote diseases. A slash denotes structurally equivalent cases collapsed into a single row; e.g., BE/BF indicates results for cases BE and BF combined.

Robust attenuation of learning about the blocked cue is shown by the test case BE/BF. People preferred Disease 4, associated with the control symptoms E and F, over Disease 2, associated with the blocked symptom B, 58.1% to 22.5%, χ^2 (df = 1, N = 129)/4 = 6.30, p < .05. (Again, the χ^2 value was divided by the number of repetitions seen by each participant, as a very conservative precaution against a possible lack of independence.) Various other statistical analyses were presented in Kruschke and Blair (2000).

These results are inconsistent with the hypothesis that blocking is caused entirely by lack of learning about the blocked cue. Instead, the results are consistent with the hypothesis that blocking is caused, at least in part, by learned inattention. The model fits presented below will support the attentional hypothesis, in that models with attentional shifting fit the data well, but constrained models without attentional shifting fail to fit the data.

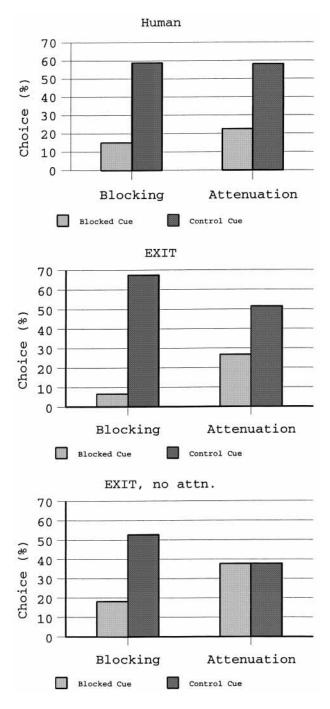


FIG. 2. Essential data (shown in boldface font in Tables 2, 3, and 4) from the experiment examining learning after blocking, with predicted values from the models, are shown. The left-hand bars, labeled "Blocking," plot data from the first test-phase item BH/BI. In this case the "blocked cue" is B, with the corresponding choice of disease 1, and the "control cue" is H or I, with the corresponding choice of disease 6. The right-hand bars, labeled "Attenuation," plot the data from the second test-phase item BE/BF. In this case the "blocked cue" is B, with the corresponding choice of disease 2, and the "control cue" is E or F, with the corresponding choice of disease 4. The experiment design appears in Table 1.

Fit of EXIT to Blocking Data

EXIT was trained on the same 40 sequences that were experienced by the 40 human participants, and the model's mean choice probabilities were fit to the 132 percentages shown in Table 2. The rows of the table are constrained to sum to 1.0, so there are 110 degrees of freedom in the data. The discrepancy between predicted and empirical proportions was assessed using the log-likelihood measure, $G^2 = 2N \sum_i f_i \log(f_i/m_i)$, where N is the number of subjects, f_i is the human choice proportion, m_i is the model's predicted proportion, and the index i runs over all 132 data points. (The values in the formula for G^2 are proportions, which are multiplied by 100 to get the percentages displayed in the tables.) The parameter values were searched using a hill-climbing method, which was started from several widely different points so that the fits reported below are likely to be globally optimal.

When N is large and when the predicted proportions are not too extreme, then G^2 is distributed approximately as χ^2 , and therefore goodness of fit can be tested. Many of the predicted proportions in the current fits do approach extreme values, however, and so the distribution of G^2 is not necessarily accurately approximated by χ^2 and we cannot assess the goodness of fit from the critical values of χ^2 . Nevertheless, G^2 remains useful as a descriptive measure of discrepancy between data and predictions, and G^2 weighs discrepancies from extreme proportions more heavily than discrepancies from moderate proportions.

EXIT fits the data fairly well, with $G^2(104) = 73.13$, for the parameter values c = 0.348, P = 1.07, $\phi = 4.43$, $\lambda_g = 1.27$, $\lambda_w = 0.316$, and $\lambda_x = 0.0121$. For these parameter values, the RMSD is 0.0341, but this was not minimized.

The second column of Table 3 shows the value of G^2 for each test case individually. These values sum to the total G^2 that was minimized by the parameter search. As each row has about 5 degrees of freedom, a row G^2 of about 10 or higher indicates an item that is poorly fit by the model. This criterion for G^2 is only of heuristic value, because G^2 might not be well approximated by χ^2 in these cases.

The only item for which the model shows a poor fit is the case AD, presumably because the model has no mechanism for strategic guessing. The problem is that in the first test phase the model gives strong choice preferences only to diseases that it has actually been trained on prior to the test phase; i.e., diseases 1, 3, and 6. People, to the contrary, select other, theretofore unseen diseases (2, 4, and 5) notably often. This was described earlier as the result of strategic guessing, whereby people might reason, "this is a case I haven't seen before, therefore it must be a disease I haven't seen before" (Kruschke & Bradley, 1995; Kruschke & Erickson, 1995). Strategic guessing might be especially strong for the case AD, because both individual symptoms have been learned thoroughly in prior training, and so it is obvious to people that this combination has not been seen before. The model has no mechanism for strategic guessing, and therefore cannot account for people's stronger selection of diseases 2, 4, and 5. Even for this case AD, however, the model nicely shows the human preference for disease 1 over disease 3.

EXIT robustly shows blocking and attenuation of learning after blocking. The critical test cases are shown in boldface font in Table 3 and are graphed in Fig. 2.

TABLE 3

Best Fit by EXIT to Choice Percentage in Table 2

	Disease						
Symptoms	G^2	1	2	3	4	5	6
		7	Γest for Bloc	king			
вн/ві	4.82	6.6	6.5	6.5	6.5	6.5	67.4
AB	0.36	80.3	3.9	3.9	3.9	3.9	3.9
D	3.22	1.1	1.1	94.4	1.1	1.1	1.1
HI	2.38	1.2	1.2	1.2	1.2	1.2	93.8
BD	5.46	5.8	5.7	71.2	5.7	5.7	5.7
AH/AI	3.51	54.1	3.9	3.9	3.9	3.9	30.2
AD	14.29	52.0	3.7	33.1	3.7	3.7	3.7
DH/DI	3.17	4.4	4.4	44.4	4.4	4.4	38.1
		Те	st for Atten	uation			
BE/BF	4.56	5.4	26.8	5.4	51.5	5.4	5.4
A	0.13	94.4	1.1	1.1	1.1	1.1	1.1
ABC	2.23	13.7	72.4	3.5	3.5	3.5	3.5
D	1.97	1.1	1.1	94.4	1.1	1.1	1.1
DEF	1.89	3.3	3.3	12.4	74.2	3.3	3.3
G	1.31	1.1	1.1	1.1	1.1	94.3	1.1
GHI	3.00	2.6	2.6	2.6	2.6	8.1	81.4
CE/CF	4.09	4.4	47.8	4.4	34.5	4.4	4.4
BH/BI	4.50	4.8	23.4	4.8	4.8	4.3	57.9
CH/CI	1.66	4.1	44.3	4.1	4.1	3.8	39.6
AB	3.30	53.3	26.9	4.9	4.9	4.9	4.9
AC	2.46	28.6	55.8	3.9	3.9	3.9	3.9
DE/DF	2.28	4.5	4.5	34.4	47.4	4.5	4.5
GH/GI	2.54	4.2	4.2	4.2	4.2	27.8	55.5

Note. Notation is the same as in Table 2.

In the test for blocking, i.e., the case BH/BI, EXIT strongly prefers the control-symptom disease over the blocked-symptom disease. In the test for attenuation after blocking, EXIT again strongly prefers the control-symptom disease over the blocked-symptom disease.

EXIT shows attenuation of learning about the blocked cue because of attentional shifting and learning. In the first phase of training (see Table 1), a strong association between cue A and outcome 1 is learned. In the initial trials of the second phase, when presented with cases of $AB \rightarrow 1$, attention is spread over both cues, which implies that cue A is not getting as much attention as it did in the first phase (when cue A was presented by itself). This lower attention to cue A causes the predicted activation of outcome 1 to be lower than it would be if cue A were presented alone. The first response to this error is to shift attention toward cue A, away from cue B. This shift is then learned, so that in subsequent presentations of cues A or B attention is directed more to cue A than to cue B. In particular, when the case

 $ABC \rightarrow 2$ is presented in the third phase of training, attention is initially (before corrective feedback appears) directed away from cue B, with cue C getting more attention than cue B. When the corrective feedback is presented, attention shifts away from cue A, but the initial disadvantage of cue B persists, leaving learning about cue B attenuated.

Fit by EXIT with No Attention Shifting. When the attentional shifting in EXIT is turned off (i.e., when $\lambda_g = 0$, $\lambda_\alpha = 0$, and c is irrelevant), the model becomes a form of the Rescorla–Wagner model, with two extra qualities: EXIT retains (1) the attentional capacity parameter and (2) the Luce/softmax rule for mapping output activations to choice probabilities. Of course, when the attentional capacity power P is very large (and when all inputs have the same salience), it is tantamount to no capacity limitation at all, as in the Rescorla–Wagner model. Therefore, it was anticipated that this version of the model, like the Rescorla–Wagner model, should show blocking, but no attenuation of learning after blocking.

TABLE 4

Best Fit by EXIT with No Attention Shifting to Choice Percentages in Table 2

	Disease						
Symptoms	G^2	1	2	3	4	5	6
		7	Γest for Bloc	king			
вн/ві	1.82	18.1	7.3	7.3	7.3	7.3	52.7
AB	6.05	92.8	1.4	1.4	1.4	1.4	1.4
D	3.36	1.2	1.2	93.9	1.2	1.2	1.2
HI	3.12	1.8	1.8	1.8	1.8	1.8	91.2
BD	1.37	14.2	5.7	62.8	5.7	5.7	5.7
AH/AI	2.07	69.9	2.7	2.7	2.7	2.7	19.3
AD	21.94	63.5	2.4	26.7	2.4	2.4	2.4
DH/DI	2.93	4.5	4.5	49.4	4.5	4.5	32.5
		Te	est for Atten	uation			
BE/BF	10.19	6.7	37.7	3.7	37.7	7.1	7.1
A	1.28	90.0	3.5	1.6	1.6	1.6	1.6
ABC	3.21	11.3	69.3	4.8	4.8	4.8	4.8
D	2.83	1.6	1.6	90.3	3.4	1.6	1.6
DEF	2.66	4.9	4.9	9.5	70.7	4.9	4.9
G	2.95	1.9	1.9	1.9	1.9	89.8	2.5
GHI	4.85	2.1	2.1	2.1	2.1	4.6	87.1
CE/CF	2.02	2.8	39.3	3.8	39.3	7.4	7.4
BH/BI	4.08	4.3	24.2	4.5	4.5	2.8	59.6
CH/CI	12.33	1.8	24.9	4.7	4.7	2.9	61.2
AB	6.66	41.7	39.1	4.8	4.8	4.8	4.8
AC	8.65	22.4	52.0	6.4	6.4	6.4	6.4
DE/DF	3.03	5.9	5.9	28.5	47.9	5.9	5.9
GH/GI	5.40	4.1	4.1	4.1	4.1	20.9	62.7

Note. Notation is the same as in Table 2.

The best fit of EXIT with no attention-shifting produced $G^2(107) = 112.78$, with parameter values of P = 1.16, $\phi = 4.35$, and $\lambda_w = 0.186$ (and a corresponding RMSD of 0.0565). The increase in G^2 of 39.65, for three degrees of freedom, is highly significant, so full EXIT fits much better than EXIT without attention. The predictions of the constrained model are shown in Table 4. As anticipated, the model nicely shows blocking for the test case BH/BI in the first test phase, but does not show any attenuation after blocking, for case BE/BF, in the final test phase. Specifically, for case BE/BF, the model without attention shows equal preference (37.7 %) for diseases 2 and 4, unlike people.

Interim Summary and Preview

To this point in the article, the EXIT model has been described and successfully fit to data that demonstrate attenuated learning about a blocked cue. EXIT is able to fit the data because of its attentional shifting and learning. Associative weight learning by itself, as in the Rescorla–Wagner model, cannot show attenuation after blocking. Several other experiments in my lab have shown robust attenuation of learning about a blocked cue. Some of these other experiments are reported in Kruschke and Blair (2000) and Kruschke (in preparation).

Later in the article it will be shown how a special case of EXIT is very similar to the model of attentional changes in animal learning proposed by Mackintosh (1975). Other applications of EXIT and related models to human data will also be summarized later.

In the next section, a different connectionist implementation of attentional shifting is described and fit to the blocking data. This implementation, in a mixture of experts architecture, will also be shown to have a special case that is very similar to the model proposed by Mackintosh (1975).

MIXTURE OF EXPERTS MODEL

In the recent connectionist modeling literature, a recurring question is how to automatically decompose complex learning problems into subproblems that are individually more easily solved than the overall problem. One approach to this issue is the *mixture of experts* framework introduced by Jacobs *et al.* (1991). The underlying notion is that a connectionist network can have several subnetworks, also called modules, each of which can discover and learn a subproblem. Each of these modules becomes an "expert" for its particular subproblem. The output and learning of these expert modules are controlled by another module called the "gating" network, which learns how to allocate the experts in specific situations. Thus, the scheme as a whole is a "mixture of experts." The mixture of experts framework has been used extensively in engineering applications and is gaining popularity as a framework for models in psychology (for a recent review see Jacobs, 1997).

The gating network learns to allocate *attention* to whichever module is doing the best job of prediction. This allocation is stimulus specific, so that attention can be

directed to one expert module for one stimulus, but to a different expert module for a different stimulus. This ability to allocate attention is the primary motivation for considering the mixture of experts approach in this article.

Attention affects both the output of the model and its learning. Thus, the expert module that is getting the most attention has the most influence on the overall output of the model. The module with the most attention also gets a much stronger error signal than the other modules, and so it learns more about the current stimulus than the other modules do. In this way the different modules can learn different aspects of the problem.

A mixture of experts approach was previously used to model the learning of rules and exceptions in classification (Erickson & Kruschke, 1998; Kruschke & Erickson, 1994). One expert module consisted of exemplars, whereas other expert modules instantiated rules. The model, called ATRIUM, exhibited trends in both learning and generalization much like people. Unlike the EXIT model, however, the shifts of attention in ATRIUM were relatively gradual and executed at the same time as the changes in associative weights. The new mixture of experts model described in the present article extends this previous work by allowing the attention shift between modules to be relatively *rapid* and to occur *before* associative weight changes, just as was done in EXIT (and in RASHNL, see Kruschke & Johansen, 1999).

In the present application, each individual cue acts as a distinct expert, trying to predict the correct output. The gating module learns which individual cues are efficacious for particular stimuli. This arrangement can account for blocking, at least in principle, because the gating module learns to ignore the expert module that contains the blocked cue.

In this section, a variant of Jacobs *et al.*'s (1991) approach is described and fit to the data from the attenuation-after-blocking experiment summarized earlier in the article. The original mixture of experts approach is briefly described later, and both the original version and its variant are shown to imply attentional shifts and associative weight changes virtually identical with those proposed by Mackintosh (1975) in his model of animal learning.

Activation Propagation in the Expert Modules

As mentioned above, each component cue comprises an "expert" that attempts to individually predict the outcome. As shown in Fig. 3, the *i*th expert module consists of a single cue node that represents the presence or absence of the *i*th cue, connected to a full set of output nodes. For the module *i*, the activation of the *k*th output node is given by

$$a_{ki}^{\text{out}} = w_{ki} a_i^{\text{in}}. \tag{11}$$

Note that the output node activations within the *i*th module are affected only by the *i*th cue. Note also that attention has no role *within* each expert module; attention does *not* gate the input activations.

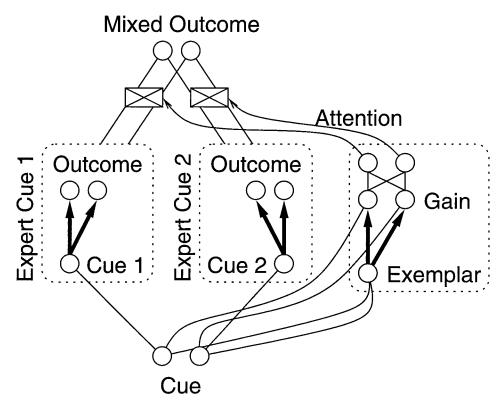


FIG. 3. Architecture for the mixture of experts model is shown. Thick arrows denote learnable associative weights. Inside the attention module, the crossing lines from gain nodes to attention nodes represent the normalization function in Eq. (5). The X's in boxes on the output lines represent the multiplicative weighting of the attention on the modular outputs, as expressed in Eq. (15). Note that the attentional module is the same as the attentional system in the EXIT model, diagrammed in Fig. 1.

Choice Probabilities. The choice probabilities are computed within each expert module according to the Luce/softmax rule (Eq. (2)). The probability that expert i selects category c is given by

$$p_i(c) = \exp(\phi a_{ci}^{\text{out}}) / \sum_k \exp(\phi a_{ki}^{\text{out}}).$$
 (12)

These modular probability distributions are combined multiplicatively to generate an overall choice probability, as follows:

$$p(c) = \prod_{i} p_{i}(c)^{\alpha_{i}} / \sum_{k} \prod_{i} p_{i}(k)^{\alpha_{i}}$$
(13)

Note that attention plays a role by causing the more strongly attended-to modules to have greater influence in the overall choice. This attentionally weighted, multiplicative combination of probability distributions ensures that modules that are merely guessing have no influence on the choice. That is, if a module has choice probabilities of just 1/K, where K is the number of categories, this factor cancels out in Eq. (13) and the module has no influence. Massaro (1987, Sect. 8.10.3) has

argued that this multiplicative combination of distributions can be more optimal than an additive combination. Unlike Erickson and Kruschke (1998), the overall prediction of the model is *not* given by $p(c) = \sum_{i} \alpha_{i} p_{i}(c)$.

Another important quality of the multiplicative combination of modular probabilities is that it turns out to be identical with the probability mapping used in EXIT. That is, substituting the definitions of modular output activation from Eq. (11) and of modular probability from Eq. (12) into the definition of mixed probability in Eq. (13) yields, after some algebraic simplification,

$$p(c) = \exp\left(\phi \sum_{i} w_{ci} \alpha_{i} a_{i}^{\text{in}}\right) / \sum_{k} \exp\left(\phi \sum_{i} w_{ki} \alpha_{i} a_{i}^{\text{in}}\right).$$

This is the same as what is obtained by substituting the definition of summed output activation from EXIT's Eq. (1) into EXIT's probability mapping, Eq. (2). Therefore, any behavioral differences between the models cannot be attributed to the mapping from output activations to choice probabilities. What differs between the models is the precise manner by which the weights and attention strengths are learned and used, as described below.

Base Rates. Category base rates are learned and attended to in the mixture of experts in a manner directly analogous to the way base rates are handled in EXIT. In a base-rate expert module, a bias cue, representing the response prompt and activated on every trial, is connected to all the outcome nodes, and this module competes with the other modules. Appendix 2 shows how this is essentially equivalent to the method proposed in the original ADIT model (Kruschke, 1996a).

Activation Propagation in the Attentional Gating Module

The attentional gating module for the mixture of experts is implemented here exactly as described for EXIT, except that the gains have a capacity (normalization power) fixed at P = 1 in Eq. (5). This is because the original mixture of experts approach (Jacobs *et al.*, 1991) treated each attention strength as the *probability* that the corresponding module is selected to apply to the current stimulus, to the exclusion of the other modules. Therefore the attention strengths must sum to unity, and this is achieved by setting the capacity to unity.

Attention Shifts

The goal of the mixture of experts model is to minimize error. The first step in minimizing error is shifting attention among the modules according to gradient descent on overall error. The error of a module is defined as the standard sumsquared error, previously used in numerous applications of back-propagation in connectionist networks (Rumelhart *et al.*, 1986, 1995). Thus, the error generated by the *i*th module is defined as

$$E_i = .5 \sum_{k} (t_{ki} - a_{ki}^{\text{out}})^2, \tag{14}$$

and the overall, mixed error of the network is defined as

$$E = \sum_{i} \alpha_{i} E_{i}. \tag{15}$$

Note the attentional weighting of the modular errors in the mixed error (Eq. (15)).

The modular gains are adjusted by gradient descent to reduce overall error, which yields the following:

$$\Delta g_{I} = -\lambda_{g} \frac{\partial E}{\partial g_{I}}$$

$$= -\lambda_{g} \sum_{i} E_{i} \frac{\partial \alpha_{i}}{\partial g_{I}}$$

$$= -\lambda_{g} \sum_{i} E_{i} \frac{1}{\sum_{j} g_{j}} (\kappa_{iI} - \alpha_{i})$$

$$= \lambda_{g} \frac{1}{\sum_{i} g_{i}} (E - E_{I}).$$
(16)

The penultimate line in Eq. (16) contains the expression for $\partial \alpha_i/\partial g_I$ derived previously in Eq. (7), with P=1. The last line of Eq. (16) shows that the attention to the *I*th module increases if and only if its error is less than the overall mixed error. In this way, attention is shifted toward those modules that are doing the best job of predicting the correct output.

Learning of Associations

After the attention is shifted, the associative weights are adjusted by gradient descent on the remaining error. For the weights in the expert modules, this yields

$$\Delta w_{ki} = -\lambda \frac{\partial E}{\partial w_{ki}}$$

$$= \lambda \alpha_i (t_{ki} - w_{ki} a_i^{\text{in}}) a_i^{\text{in}}.$$
(17)

Note that a weight changes only to the extent that its module is being attended to. As explained previously in the description of EXIT (Eq. (10)), the weights from exemplar nodes to gain nodes in the gating module are also adjusted according to gradient descent on error. The shifted attention strengths serve as the target values.

Free Parameters in the Mixture of Experts Model. The mixture of experts model has all the parameters of EXIT except one, the attention normalization power (which is fixed at P = 1). Thus, the mixture of experts model has five freely estimated parameters.

Fit by Mixture of Experts to Blocking Data

Table 5 and Fig. 4 show the predictions of the best fit obtained for the mixture of experts model. The fit resulted in $G^2(105) = 73.27$, for parameter values c = 0.378,

TABLE 5

Best Fit by Mixture of Experts Model to Choice Percentages in Table 2

	Disease						
Symptoms	G^2	1	2	3	4	5	6
		7	Γest for Bloc	king			
BH/BI	5.32	6.3	6.3	6.3	6.3	6.3	68.7
AB	0.63	77.8	4.4	4.4	4.4	4.4	4.4
D	3.06	1.0	1.0	95.0	1.0	1.0	1.0
HI	2.30	1.1	1.1	1.1	1.1	1.1	94.5
BD	5.34	6.0	6.0	70.1	6.0	6.0	6.0
AH/AI	4.93	50.6	4.2	4.2	4.2	4.2	32.4
AD	13.01	50.3	4.2	33.0	4.2	4.2	4.2
DH/DI	3.51	4.3	4.3	42.4	4.3	4.3	40.2
		Te	est for Atteni	ıation			
BE/BF	4.17	5.1	25.8	5.1	53.9	5.1	5.1
A	0.16	95.0	1.0	1.0	1.0	1.0	1.0
ABC	2.27	13.9	71.2	3.7	3.7	3.7	3.7
D	2.07	1.0	1.0	95.0	1.0	1.0	1.0
DEF	1.96	3.3	3.3	11.5	75.2	3.3	3.3
G	1.20	1.0	1.0	1.0	1.0	95.0	1.0
GHI	2.90	2.7	2.7	2.7	2.7	8.8	80.2
CE/CF	3.72	4.5	46.3	4.5	35.6	4.5	4.5
BH/BI	4.29	4.6	22.9	4.6	4.6	4.6	58.5
CH/CI	1.81	4.3	43.6	4.3	4.3	4.3	39.2
AB	3.31	57.8	23.5	4.7	4.7	4.7	4.7
AC	2.86	28.0	55.6	4.1	4.1	4.1	4.1
DE/DF	2.02	4.4	4.4	32.3	49.9	4.4	4.4
GH/GI	2.43	4.1	4.1	4.1	4.1	28.0	55.8

Note. Notation is the same as in Table 2.

 $\phi=4.56$, $\lambda_g=4.24$, $\lambda_w=0.344$, and $\lambda_x=0.0110$. These parameter values yield RMSD = 0.0347 (for proportions; i.e., 3.47 for percentages), but other parameter values yield a slightly smaller RMSD. The mixture of experts model shows good fits to the two critical test stimuli: In the test for blocking, stimulus BH/BI shows a much higher preference for disease 6 than for disease 1. In the test for attenuation, stimulus BE/BF shows a much higher preference for disease 4 than for disease 2.

The mixture of experts model shows attenuation of learning about the blocked cue because of attentional shifts and learning, in much the same way as in EXIT. In the first phase of training (see Table 1), a strong association between cue A and outcome 1 is learned. In the initial trials of the second phase, when presented with cases of $AB \rightarrow 1$, attention is spread over both cues' modules, which implies that the cue-A module is not getting as much attention as it did in the first phase (when cue A was presented by itself). This lower attention to the cue-A module causes the mixed error to be greater than it would be if cue A were presented alone. The first

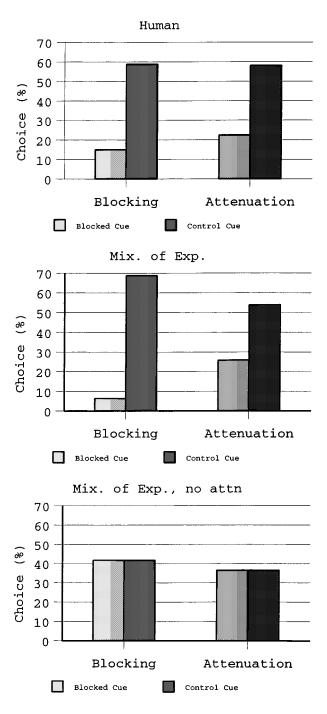


FIG. 4. Essential data (shown in boldface font in Tables 2, 5, and 6) from the experiment examining learning after blocking, with predicted values from the models is shown. The left-hand bars, labeled "Blocking," plot data from the first test-phase item BH/BI. In this case the "blocked cue" is B, with the corresponding choice of disease 1, and the "control cue" is H or I, with the corresponding choice of disease 6. The right-hand bars, labeled "Attenuation," plot data from the second test-phase item BE/BF. In this case the "blocked cue" is B, with the corresponding choice disease 2, and the "control cue" is E or F, with the corresponding choice of disease 4. The experiment design appears in Table 1.

response to this error is to shift attention toward the cue-A module, away from the cue-B module. This shift is then learned, so that in subsequent presentations of cues A or B attention is directed more to the cue-A module than to the cue-B module. In particular, when the case $ABC \rightarrow 2$ is presented in the third phase of training, learning in the cue-B module is attenuated.

Fit by Mixture of Experts with No Attention Shift. To illustrate the critical role of attentional shifting in the mixture of experts model, the attentional shift and attention learning rates were fixed at zero, and then the best possible fit to the data was found. The best fit of this constrained model yielded $G^2(108) = 145.77$, for parameter values of $\phi = 4.89$ and $\lambda_w = 0.271$ (which yielded RMSD = 0.0706). Relative to the full mixture of experts model, this increase in G^2 of 72.50, for three degrees of freedom, is highly significant. That is, the attentional model fits the data far better than the no-attention form.

TABLE 6

Best Fit by Mixture of Experts Model with No Attention Shifting to Choice Percentages in Table 2

	Disease						
Symptoms	G^2	1	2	3	4	5	6
		7	Γest for Bloc	king			
вн/ві	14.71	41.7	4.1	4.1	4.1	4.1	41.7
AB	12.32	95.9	0.8	0.8	0.8	0.8	0.8
D	2.92	0.7	0.7	96.4	0.7	0.7	0.7
HI	2.29	0.9	0.9	0.9	0.9	0.9	95.3
BD	12.62	39.4	3.9	45.0	3.9	3.9	3.9
AH/AI	8.11	45.0	3.9	3.9	3.9	3.9	39.4
AD	15.19	42.6	3.7	42.6	3.7	3.7	3.7
DH/DI	4.00	3.9	3.9	45.0	3.9	3.9	39.4
		Те	st for Atten	uation			
BE/BF	13.20	10.0	36.5	5.7	36.5	5.7	5.7
Α	3.63	85.8	6.2	2.0	2.0	2.0	2.0
ABC	3.27	19.3	65.5	3.8	3.8	3.8	3.8
D	5.29	2.0	2.0	85.5	6.4	2.0	2.0
DEF	1.94	4.0	4.0	14.0	69.9	4.0	4.0
G	5.76	2.0	2.0	2.0	2.0	85.3	6.5
GHI	3.72	2.3	2.3	2.3	2.3	7.9	83.1
CE/CF	3.60	6.0	38.1	6.0	38.1	6.0	6.0
BH/BI	8.43	7.8	28.6	4.5	4.5	4.5	50.1
CH/CI	7.36	4.6	29.6	4.6	4.6	4.6	51.8
AB	6.82	42.9	42.1	3.7	3.7	3.7	3.7
AC	3.41	30.1	51.6	4.6	4.6	4.6	4.6
DE/DF	2.04	4.6	4.6	29.8	51.9	4.6	4.6
GH/GI	5.16	3.3	3.3	3.3	3.3	21.3	65.5

Table 6 and Figure 4 show the predictions of the mixture of experts model with no attention shifting. Several cases are badly misfit. Most importantly, the model can show neither blocking nor attenuation of learning after blocking. The test for blocking, case BH/BI, shows equal preference (41.7 %) for diseases 1 and 6. The test for attenuation after blocking, case BE/BF, shows equal preference (36.5 %) for diseases 2 and 4. Thus, the full mixture of experts model relies completely on its attentional shifting mechanism to account for blocking (which was also the case in Mackintosh's (1975) approach, as will be described below).

Interim Summary and Preview

To this point in the article, two different connectionist models have been described that implement the ideas of rapidly shifted and learned attention. The models differ in the locus of attentional control: In EXIT, attention modulates the input activations; in the mixture of experts, attention modulates the output activations (of the expert modules). EXIT extends ADIT (Kruschke, 1996a) by *learning* the shifted distributions of attention. The mixture of experts model extends ATRIUM (Erickson & Kruschke, 1998) by allowing *rapid* attentional shifts (and by using a multiplicative combination of probabilities instead of an additive combination).

Both models were fit to data that demonstrate attenuation of learning about a blocked cue. When attentional shifting and learning are shut off, neither model can exhibit attenuation of learning about a blocked cue. In the mixture of experts model, blocking is also entirely the result of attentional shifting. In EXIT, blocking is the result of both attentional shifting and associative learning. This difference between the models stems from the different error signals used to drive associative weight changes. In EXIT the error is computed at each category node as the desired value minus the *summed* prediction from all the input cues (Eq. (9)), whereas in the mixture of experts model the error is computed at each category node as the desired value minus the *unique* prediction from the individual expert cue (Eq. (17)).

The next sections relate these two models of human category learning to Mackintosh's (1975) classic model of animal learning.

4. THE ATTENTIONAL THEORY OF MACKINTOSH (1975)

Theoretical Motivations for Mackintosh's Formalization

The purpose of this section is to summarize the four main ideas that Mackintosh formalized. It will be noted that not all of the motivations are equally countenanced; some are expressions of theoretical parsimony, whereas other motivations are central commitments. It is important to make these motivations explicit because the special cases of the connectionist models will contain slightly different details, and it will be argued that the differences do not violate the central motivations of Mackintosh's model.

The primary principle to be formalized was that the attention devoted to a cue should change to reflect the relevance of that cue for predicting the reinforcement:

"... the probability of attending to relevant stimuli typically increases, while the probability of attending to irrelevant stimuli typically decreases" (Mackintosh, 1975, p. 278).

Mackintosh (1975) also argued that changes in associability and in associative strength are influenced by how well each cue *uniquely* predicts the reinforcement, not by how well the aggregate of all present stimuli predicts the reinforcement. Unlike the Rescorla–Wagner (1972) model, in which predictiveness is determined by the *summed* influence of all cues present, Mackintosh (1975) suggested that learning attempts to maximize the extent to which each cue individually accounts for the reinforcement. This implies rather different mechanisms at work in explaining classic phenomena such as blocking (Kamin, 1969). As Pearce and Hall (1980, p. 536) noted, "In blocking and related phenomena, the whole of the explanatory burden is borne by the mechanisms dealing with changes in [cue associability]."

A third motivation was that the mechanism for updating attention should not explicitly rely on the *inverse hypothesis*, which posits that increased attention to one cue necessarily causes decreased attention to other cues. Mackintosh did not assert that this hypothesis was wrong, but argued that it was not absolutely necessary for explaining the empirical data of interest and that therefore it should not be presumed.

A fourth motivation was the idea that the attention that modulates the associability of a cue need not necessarily be the same as the attention that modulates the response control of the cue. Thus, it is possible that a cue could have a high associability, yet not have a large influence on responding. Conversely, it is possible for a cue to strongly influence responding, yet have a low associability. Mackintosh (1975) was not deeply committed to this separation and even stated, "It may be necessary to assume that although such a stimulus maintains its associative strength, a decline in [associability] will decrease the probability that it will control responding" (p. 294).

Mackintosh's (1975) Formal Model

The purpose of this section is to summarize Mackintosh's formal model in contemporary connectionist notation, while being as faithful as possible to the original details. Use of this notation permits consistency of expression with the connectionist models described earlier. In this section I will also discuss how each of the four motivating theoretical principles is realized (or not realized) in the formal models. This in turn will be useful for evaluating the differences between the original model and its connectionist generalizations.

Mackintosh (1975) was not explicit with regard to how predictions from individual cues were combined into an overall response level for compound stimuli. I shall simply assume that the overall response is given by

$$a^{\text{out}} = \sum_{i} w_i a_i^{\text{in}} . \tag{18}$$

This is a reasonable assumption because of Mackintosh's reference to the Rescorla–Wagner model. Note that this expression does not include any terms for attention; that is, attention plays no role in direct control of responses. This implements the fourth motivation above, whereby attention in response generation is not necessarily the same as attention in learning. Recall, however, that Mackintosh was not committed to the absence of attention from response production, he merely did not want to be forced to posit it. If attention were to play a role in controlling response output, it would reasonably appear in Eq. (18) as cue-specific multipliers α_i on each cue activation. In this case, the output activation could be expressed as

$$a^{\text{out}} = \sum_{i} w_i \alpha_i a_i^{\text{in}} \,. \tag{19}$$

The associative weight from a cue is adjusted to reduce the difference between the outcome predicted by that *individual* cue and the actual outcome. This weight change is formally specified as

$$\Delta w_I = \lambda \alpha_I (t - w_I a_I^{\text{in}}) a_I^{\text{in}}, \tag{20}$$

where λ is a general learning rate parameter that applies equally to all cues and α_I is a cue-specific learning rate which can differ from one cue to another. Both λ and α_I are assumed to lie between zero and one. This Eq. (20) corresponds to Eq. (2) from Mackintosh (1975). Note that the α_I in Eq. (20) expresses the associability of cue I, whereas the α_I in Eq. (19) expresses the response-evoking power of cue I.

Equation (20) is an expression of Mackintosh's second motivation described above, i.e., that changes in association should depend on how well each cue *uniquely*, or individually, predicts the outcome. The weight change formula implements this idea because the weight change is proportional to the difference between the teacher and the weighted activation of the single, unique cue. This can be contrasted with the Rescorla–Wagner (1972) weight-change rule, wherein a weight was adjusted proportionally to the difference between the teacher and *the sum of all weighted inputs*; i.e.,

$$\Delta w_I = \lambda \alpha_I \left(t - \sum_i w_i a_i^{\text{in}} \right) a_I^{\text{in}}. \tag{21}$$

This contrast between unique and summed inputs will recur below.

With regard to changes in the associability α_I of cue I, Mackintosh suggested that:

The intuition that we require to formalize is that α_I should increase if I predicts an otherwise unexpected reinforcer, while α_I should decrease if I signals no change in reinforcement from the level expected on the basis of other events. There are, presumably, a number of ways in which this might be done, but possibly the simplest is as follows. (Mackintosh, 1975, p. 287, with subscripts changed from A to I)

In modified notation, Mackintosh's formalization can be expressed as

$$\Delta \alpha_I > 0 \quad \text{if} \quad |t - w_I a_I^{\text{in}}| < \left| t - \sum_{i \neq I} w_i a_i^{\text{in}} \right|, \tag{22}$$

$$\Delta \alpha_I < 0 \qquad \text{if} \quad |t - w_I a_I^{\text{in}}| > \left| t - \sum_{i \neq I} w_i a_i^{\text{in}} \right|. \tag{23}$$

(These Eqs. (22) and (23) correspond to Eqs. (4) and (5), respectively, of Mackintosh, 1975.) An important aspect not explicitly notated in these equations is that the change in attention is supposed to be effective on the subsequent trial with the same stimulus, not on the current trial.

Equation 23 (above) was modified from the original in one critical way. The original expression used the relation " \geqslant " instead of ">." Mackintosh motivated the original usage by saying "... it is only the phenomenon of latent inhibition that necessitates the otherwise rather unhappy assumption that α_I declines even when $|t-w_Ia_I^{\rm in}|$ is equal to $|t-\sum_{i\neq I}w_ia_i^{\rm in}|$. For reasons of theoretical symmetry, if for no others, one would expect this equality to produce no change in α_I " (Mackintosh, 1975, p. 289, with mathematical notation changed). In the generalizations of the model in terms of the connectionist networks, the asymmetry is jettisoned. A novel theoretical treatment of the pre-exposure effect (latent inhibition) is described in a subsequent section of this article.

These formalizations of the conditions for attention change possibly constitute an inconsistency with Mackintosh's motivating principles, in so far as the formalisms embrace the "summed effects" approach on their right-hand side. Indeed, Mackintosh stated that "These rules [Eqs. (22) and (23), above] are certainly not to be thought of as final, but rather as illustrating that it is possible to express this informal idea reasonably precisely" (Mackintosh, 1975, p. 290). If summed effects are rejected in favor of unique (individual cue) effects, then expressions of Mackintosh's intuition, quoted above, might instead be

$$\Delta \alpha_I > 0$$
 if $|t - w_I a_I^{\text{in}}| < \frac{1}{N-1} \sum_{i \neq I} |t - w_i a_i^{\text{in}}|,$ (24)

$$\Delta \alpha_I < 0$$
 if $|t - w_I a_I^{\text{in}}| > \frac{1}{N - 1} \sum_{i \neq I} |t - w_i a_i^{\text{in}}|.$ (25)

In other words, the attention to cue *I* should increase if and only if the prediction it uniquely generates is more accurate than the average unique prediction of the other cues.

Summary and Preview

Mackintosh stated a rule for changing associative weights based on unique, cuespecific predictions and cue-specific associabilities (Eq. (20)). This rule for weight changes was contrasted with the summed-effects incorporated into the Rescorla–Wagner weight change rule (Eq. (21)).

Mackintosh stated a rule for changing cue-specific associabilities (Eqs. (22) and (23)) based on the cue's predictive accuracy relative to the predictive accuracy of the other cues present. The original rule involved an asymmetry whose only purpose was to address the pre-exposure effect. The asymmetry is jettisoned in the present treatment, with an alternative account of the pre-exposure effect introduced later.

To better reflect Mackintosh's principled rejection of summed effects in favor of unique effects, I proposed a modified formal expression for associability change in Eqs (24) and (25) that, I believe, remains faithful to Mackintosh's explicitly stated motivating intuition.

In the following sections of the article, it will be shown that the connectionist models imply essentially the same formulas as Mackintosh's when there is a single output category (corresponding to the presence or absence of the unconditioned stimulus). The only difference is that the connectionist models assume that attention plays a role in response performance as well as in learning; Mackintosh did not view the separation of attention for response and attention for learning as crucial. The EXIT model uses summed effects and corresponds with the Rescorla–Wagner formula for weight changes and Mackintosh's (1975) summed-effect formulas for attention change. The mixture of experts model uses unique effects and corresponds with Mackintosh's (1975) formula for weight changes but with the unique-effects formulas for attention change.

Mackintosh's Attentional Model as a Version of EXIT

Suppose there is just a single output category and the attentional capacity (metric power) is set to unity, i.e., P = 1. In this case, the rule for changing attention in EXIT becomes nearly identical to the rule proposed by Mackintosh. To begin, note first that when there is just a single output node and when P = 1, EXIT's formula for attention change, Eq. (8), becomes

$$\Delta g_I = (t - a^{\text{out}})(w_I a_I^{\text{in}} - a^{\text{out}}) / \left(\sum_j g_j\right)$$

Note next that the term $w_I a_I^{\rm in} - a^{\rm out}$ can be algebraically re-expressed as $(1 - \alpha_I)$ $(w_I a_I^{\rm in} - (1/(1 - \alpha_I)) \sum_{i \neq I} w_i \alpha_i a_i^{\rm in})$. Therefore the formula for attention change in EXIT becomes

$$\Delta g_I = (t - a^{\text{out}})(1 - \alpha_I) \left(w_I a_I^{\text{in}} - \frac{1}{1 - \alpha_I} \sum_{i \neq I} w_i \alpha_i a_i^{\text{in}} \right) / \left(\sum_j g_j \right). \tag{26}$$

Eq. (26) implies that

$$\Delta g_I > 0 \qquad \text{if} \quad (t - a^{\text{out}})(1 - \alpha_I) \left(w_I a_I^{\text{in}} - \frac{1}{1 - \alpha_I} \sum_{i \neq I} w_i \alpha_i a_i^{\text{in}} \right) \middle/ \left(\sum_j g_j \right) > 0. \tag{27}$$

Because $(\sum_j g_j) > 0$ and $(1 - \alpha_I) > 0$, the right-hand inequality of Eq. (27) is true if and only if

$$w_I a_I^{\text{in}} - \frac{1}{1 - \alpha_I} \sum_{i \neq I} w_i \alpha_i a_i^{\text{in}} > 0$$
 for $t - a^{\text{out}} > 0$

or

$$w_I a_I^{\text{in}} - \frac{1}{1 - \alpha_I} \sum_{i \neq I} w_i \alpha_i a_i^{\text{in}} < 0$$
 for $t - a^{\text{out}} < 0$.

These conditions in turn become the following:

$$\Delta g_I > 0$$

if

$$t - w_I a_I^{\text{in}} < t - \frac{1}{1 - \alpha_I} \sum_{i \neq I} w_i \alpha_i a_i^{\text{in}} \qquad \text{for} \quad t - a^{\text{out}} > 0$$

or (28)

$$t - w_I a_I^{\text{in}} > t - \frac{1}{1 - \alpha_I} \sum_{i \neq I} w_i \alpha_i a_i^{\text{in}} \qquad \text{for} \quad t - a^{\text{out}} < 0.$$

A directly analogous series of computations shows similarly that

$$\Delta g_I < 0$$

if

or

$$t - w_I a_I^{\text{in}} > t - \frac{1}{1 - \alpha_I} \sum_{i \neq I} w_i \alpha_i a_i^{\text{in}} \qquad \text{for} \quad t - a^{\text{out}} > 0$$
(29)

. 1 _ .

$$t - w_I a_I^{\text{in}} < t - \frac{1}{1 - \alpha_I} \sum_{i \neq I} w_i \alpha_i a_i^{\text{in}} \qquad \text{for} \quad t - a^{\text{out}} < 0.$$

Table 7 compares the formulas used in EXIT with those used by the summed-effects version of Mackintosh's model. One obvious difference between the models is that EXIT employs the attention strengths in computing the output activations, but Mackintosh's model does not. That is, everywhere the term $\sum_i w_i a_i^{\text{in}}$ appears in Mackintosh's formulas, the term $\sum_i w_i \alpha_i^{\text{in}}$ appears in EXIT's formulas (see Table 7). As said before, this use of attention in generating responses does not violate a critical principle in Mackintosh's approach.

Another difference between the models is that Mackintosh's formulas use the absolute value of the differences, but the formulas in EXIT do not, instead conditionalizing on the sign of $t-a^{\rm out}$. This implies that there are some circumstances, very rare in practice, in which the gain in EXIT and the attention in Mackintosh's

TABLE 7

Summary of Formal Expressions for Changing Selected Variables in EXIT and in Mackintosh (1975)

	Mod	Model					
Variable	EXIT, single output, $P = 1$	Mackintosh (1975), with summed effects					
$\Delta w_I =$	$\lambda \alpha_{I} \left(t - \sum_{i} w_{i} \alpha_{i} a_{i}^{\text{in}} \right) a_{I}^{\text{in}}$ (Eq. (9))	$\lambda \alpha_I \left(t - \sum_i w_i a_i^{\text{in}} \right) a_I^{\text{in}}$ (Eq. (21))					
$\Delta g_I > 0$ if	$(t - w_I a_I^{\text{in}}) < \left(t - \frac{1}{1 - \alpha_I} \sum_{i \neq I} w_i \alpha_I a_i^{\text{in}}\right)$ for $t - a^{\text{out}} > 0$ (Eq. (28))	$\begin{aligned} t-w_I a_I^{\mathrm{in}} < \left t - \sum_{i \neq I} w_i a_i^{\mathrm{ir}} \right \end{aligned}$ (Eq. (22))					
$\Delta g_I < 0$ if	$(t - w_I a_I^{\text{in}}) > \left(t - \frac{1}{1 - \alpha_I} \sum_{i \neq I} w_i \alpha_i a_i^{\text{in}}\right)$ for $t - a^{\text{out}} > 0$ (Eq. (29))	$\begin{aligned} t - w_I a_I^{\text{in}} > & \left t - \sum_{i \neq I} w_i a_i^{\text{ir}} \right \\ & (\text{Eq. (23)}) \end{aligned}$					

model change in opposite directions. For example, consider the conditions in Mackintosh's formulas for attention increase (Eqs. (22) and (23)). Suppose that after some degree of learning it is the case for a given trial that $t - \sum_{i \neq I} w_i a_i^{\text{in}} = 0.1$ and $t - w_I a_I^{\text{in}} = -0.2$ (whether this situation could ever actually arise in a real learning situation is an open question). According to Mackintosh's model, attention to the *I*th cue should *decrease*. But according to EXIT, the analogous conditions including attention multipliers (Eq. (28)) would cause an *increase* in the attentional gain on the *I*th cue. Which of these (increase or decrease) is the "correct" way to go? The direction dictated by EXIT is correct in the sense that it reduces error in the context of limited capacity attention. On the other hand, this might not correctly capture the intuition intended by Mackintosh. In any case, the similarities between the models are striking.

Mackintosh's Attentional Model as a Version of Mixture of Experts

Equation (16) implies that

$$\Delta g_I > 0$$
 if $E_I < E$. (30)

That is, the gain on the *I*th expert increases if and only if the error the expert generates is less than the overall mixed error. For purposes of mapping this result onto Mackintosh's formula for attention change, note that the right-hand side of Eq. (30) can be re-expressed as

$$\begin{split} E_I &< E, \\ E_I &< \sum_{i \neq I} \; \alpha_i E_i + \alpha_I E_I, \\ (1 - \alpha_I) \; E_I &< \sum_{i \neq I} \; \alpha_i E_i, \\ E_I &< \frac{1}{1 - \alpha_I} \; \sum_{i \neq I} \; \alpha_i E_i. \end{split}$$

Substituting this into Eq. (30) yields

$$\Delta g_I > 0$$
 if $E_I < \frac{1}{1 - \alpha_I} \sum_{i \neq I} \alpha_i E_i$. (31)

In other words, the gain on the *I*th expert increases if and only if the error it generates is less than the mixed error of the *other* experts. This result does not depend on the definition of modular error given in Eq. (14).

By substituting the expression for modular error from Eq. (14), and the expression for output activation from Eq. (11) into Eq. (31), we obtain a condition for changes in attention:

$$\Delta g_I > 0$$
 if $(t - w_I a_I^{\text{in}})^2 < \frac{1}{1 - \alpha_I} \sum_{i \neq I} \alpha_i (t - w_i a_i^{\text{in}})^2$. (32)

Directly analogous computations also imply that

$$\Delta g_I < 0$$
 if $(t - w_I a_I^{\text{in}})^2 > \frac{1}{1 - \alpha_I} \sum_{i \neq I} \alpha_i (t - w_i a_i^{\text{in}})^2$. (33)

Table 8 compares the formulas in the mixture of experts model with those used by Mackintosh (1975). These formulas for attention change are very nearly the same as the modified attention change rule for Mackintosh's model, as expressed in Eqs. (24) and (25). One difference between the equations is that Mackintosh used absolute values, whereas the mixture of experts uses squared values. If the modular error were redefined as $E_i = |t - w_i a_i^{\text{in}}|$, then the formula for $\Delta \alpha_I$ would involve absolute error instead of squared error, just as in Mackintosh's formulas, but then the formula for Δw_i derived from gradient descent would then no longer match Mackintosh's formula. In any case, whether the terms are rectified by squaring or by taking the absolute value makes little difference for the underlying principles in Mackintosh's formulation.

Original Version of Mixture of Experts. Although it will not be detailed here, it is straightforward to show, using steps analogous to those above, that the original formulation (Jacobs et al., 1991) of mixture of experts in terms of accuracies, as opposed to the squared error used here, also maps onto Mackintosh's approach. In the original approach, the accuracy of the *i*th module is defined as $A_i = \exp(-.5 \sum_k \left[t_k - a_{ik}^{\text{out}}\right]^2)$, and the overall mixed accuracy is defined as $A = \sum_i \alpha_i A_i$. The model shifts attention and adjusts weights by maximizing accuracy, rather than

TABLE 8

Summary of Formal Expressions for Selected Variables in the Mixture of Experts Model and in Mackintosh (1975)

Variable	Model						
	Mixture of experts, single output	Mackintosh (1975), with no summed effects					
$\Delta w_I =$	$\lambda \alpha_I (t - w_I a_I^{\text{in}}) a_I^{\text{in}}$ (Eq. (17))	$\lambda \alpha_I (t - w_I a_I^{\text{in}}) a_I^{\text{in}}$ (Eq. (20))					
$\Delta g_I > 0$ if	$(t - w_I a_I^{\text{in}})^2 < \frac{1}{1 - \alpha_I} \sum_{i \neq I} \alpha_i (t - w_i a_i^{\text{in}})^2$ (Eq. (32))	$ t - w_I a_I^{\text{in}} < \frac{1}{N-1} \sum_{i \neq I} t - w_i a_i^{\text{in}} $ (Eq. (24))					
$\Delta g_I < 0$ if	$(t - w_I a_I^{\text{in}})^2 > \frac{1}{1 - \alpha_I} \sum_{i \neq I} \alpha_i (t - w_i a_i^{\text{in}})^2$ (Eq. (33))	$ t-w_I a_I^{\text{in}} > \frac{1}{N-1} \sum_{i \neq I} t-w_i a_i^{\text{in}} $ (Eq. (25))					

by minimizing error. Also in the original approach, the modular gains are mapped to modular attention using the softmax rule (Eq. (2)) instead of the power transformation (Eq. (5)). That is, $\alpha_i = \exp(g_i)/\sum_j \exp(g_j)$. Appendix 1 discusses relations between the softmax transformation and the power transformation. Despite these differences, when the computational steps taken above are applied in a fashion directly analogous to the original formulation of the mixture of experts, the resulting conditions for increasing or decreasing attentional gain are again similar to the (no-summed effects version of the) rules proposed by Mackintosh (1975).

EXIT vs Mixture of Experts. Is EXIT or mixture of experts the better representation of Mackintosh's (1975) approach? The answer must be the mixture of experts model, because it, like Mackintosh's approach, relies entirely on attention to exhibit blocking (see Table 6 and the corresponding discussion). This reliance stems from there being no summed effects of cues for generating outcome activations. This also implies that EXIT is in some ways a more flexible model, as it is capable of exhibiting blocking without attenuation of learning after blocking. In preliminary fits of the models to other data (which are not further described here), EXIT has performed better than the mixture of experts. A systematic comparison of the models' abilities to fit data awaits future research.²

² The mixture of experts model might be unable to adequately exhibit *conditioned inhibition*, whereas EXIT can. In the conditioned inhibition procedure, subjects learn that $A \to 1$ and $C \to 1$ but $AB \to -1$ (not outcome 1). Cue B is said to show conditioned inhibition of outcome 1 if tests with combined cues BC show reduced prediction of outcome 1 relative to cue C by itself (i.e., a "summation test") and if subsequent learning of $B \to 1$ is attenuated relative to learning about a control cue, $D \to 1$ (i.e., a "retardation test"). The mixture of experts model can pass a summation test because it learns to enhance attention to cue B when presented with the pair AB. To the extent that the attentional shift learned for pair AB generalizes to the pair CB, there will be reduced prediction of the outcome. This learned attention, however, will work against any retardation of learning $B \to 1$.

A couple of other differences between the connectionist models and Mackintosh's (1975) approach are worth reiterating. First, the connectionist models shift attention on each trial *before* adjusting associative weights. In Mackintosh's (1975) approach, the attentional shift did not have any influence until the *subsequent* trial with the same stimulus. The extent to which this difference affects the fits to data can only be worked out by future research. A second difference between the connectionist models and Mackintosh's approach is that he used asymmetric conditions for increasing and decreasing attention in order to explain the preexposure effect (latent inhibition), whereas the connectionist models do not contain this asymmetry. An alternative treatment of the preexposure effect, within the connectionist framework, is presented in the next section.

THE PRE-EXPOSURE EFFECT (LATENT INHIBITION)

When a cue is consistently paired with an otherwise unpredicted outcome, people and animals learn to associate the cue with the outcome. The speed of learning is lower, however, if the cue has previously occurred without any consequence (Lubow & Moore, 1959; Lubow, 1989). This relatively slower learning of a cue that has been preexposed is called *latent inhibition* or *the preexposure effect*. This phenomenon has become one of the central effects to be addressed by theories of associative learning.

The connectionist models described above cannot account for the preexposure effect when it is conceptualized in its traditional way. In this tradition, the lack of any unpredicted outcome with the preexposed cue is perfectly predicted by a network with zero connection weights, and so there is no error generated by the preexposed cue. In so far as learning in the connectionist models is entirely driven by error, there is no learning and so preexposure causes no subsequent effects in the model.

Recall that Mackintosh's original formalism could address the preexposure effect only by the "rather unhappy assumption" (Mackintosh, 1975, p. 289) that attention to a cue will decrease even when its predictive error equals the predictive error of other cues. This assumption accounts for the preexposure effect because when a novel cue is presented with no concomitant outcome, there is zero error for all cues. Hence the (zero) predictive error of the novel cue equals the (zero) predictive error of the other cues, and attention to the novel cue decreases. This method for reduction in attention was expressed formally by the relation " \geq " in Eq. (23), which, by including equality instead of just inequality, was not symmetric with the relation "<" in Mackintosh's formula for increasing attention (Eq. (22)). The connectionist models derived above cannot retain this asymmetry because they do not have separate formulas for increasing and decreasing attention. Instead, a single formula for attentional change was derived from the goal of error reduction.

After a very brief review of previous models of the preexposure effect, I will describe how the effect might be addressed within the current connectionist framework. This survey of previous models is in no way intended to be exhaustive, nor is it intended to be a critique of these models. The point of the review is merely to establish that the new approach is indeed novel relative to the previous theories.

Moore and Stickney (1980) described an extension of Mackintosh's (1975) model, in which Mackintosh's formulation was symmetrized, and the preexposure effect was addressed by considering associations between the preexposed cue and the *context*. The extension assumes that the various cues are predicting not just the US, but also the cues themselves. During preexposure, the context cues are a better predictor of that context than the new cue is, and so attention to the new cue is reduced. This approach bears some similarity to the new approach described below. The formulation of Moore and Stickney (1980), however, retained Mackintosh's separation of formulas for association changes and attention changes and did not unify attentional learning and associative learning as common consequences of error reduction.

Wagner's (1981) SOP model (for a more recent version, see Brandon & Wagner, 1998) also accounts for the preexposure effect in terms of context associations. During preexposure, associations are learned from the context to the preexposed cue. In subsequent training in the same context, the preexposed cue is activated by the context associations, and this places the preexposed cue in a state of lower associability. There is no statement in this theory regarding how lower associability reduces error.

In the models of Lubow (1989) and of Pearce and Hall (1980), attention is driven toward the error value. During exposure to a novel stimulus without reinforcement, there is no error in these models, so the act of shifting attention toward the error value constitutes shifting the attention toward zero. In the model of Schmajuk et al. (1996), the attention given to a cue is also driven toward the overall error (referred to there as "novelty"; Schmajuk et al., 1996, Eq. (2), p. 324). In the model of McLaren, Kaye, and Mackintosh (1989), a modulator reinjects the cue's novelty into the learning rule, such that preexposed (no longer novel) cues are learned about more slowly than novel cues. In none of these models is there any formal statement that attentional shifting reduces error, in the sense of somehow making the model perform better after the attentional shift than before. For example, in the approach of Pearce and Hall (1980), attention decreases when the error decreases, but this decrease in attention does not itself decrease the error.

In the model of Frey and Sears (1978), attention to a cue is driven toward the value of its associative weight (there was only one associative weight because the model addressed learning about a single unconditioned stimulus). At the beginning of the preexposure phase the associative weight of the novel cue is zero, and so changing the value of the attention toward the value of the associative weight value constituted shifting the attention toward zero. Again, there was no explanation of how this attention change would improve performance or reduce error.

The Pre-Exposure Effect Is a Case of Attenuation after Blocking

In contrast with all of the previously summarized explanations of the preexposure effect, the account proposed here suggests that the preexposed cue does indeed cause error when it is initially presented. This error is reduced by shifting attention away from the cue, and the shift of attention is learned and perseverates into subsequent learning.

The nature of the error generated by the preexposed cue will be explained by analogy to blocking. In blocking, the first phase of training establishes a well-learned association between a cue and a correct response. In the second phase of training, the cue is presented with a novel redundant cue, but the same old response is all that is called for. Presentation of the redundant cue draws attention away from the cue that already perfectly predicted the correct response. This diversion of attention generates an error, i.e., a smaller magnitude of activation of the response. To reduce this error, attention shifts away from the redundant cue to the already-known cue. This shift is learned and causes attenuation of learning about the blocked cue in subsequent phases.

The preexposure effect has the same sequence of events, except that the first phase of training has happened *before* preexposure. Before preexposure, the learner has already acquired a variety of associations between cues and responses. Many of the cues and responses are internal, and so they may be called contextual cues and maintenance responses. Contextual cues are different than the bias cues that were described earlier for computing base rates. The bias cue is thought of essentially as the response prompt, but contextual cues permeate the background, both inside and outside the learner. One type of desired maintenance response might be general vigilance, and another might be varieties of preening. After prior learning, in the subsequent preexposure phase, the contextual cues are presented with a novel cue, but the same old maintenance response is all that is called for. Presentation of the novel cue draws attention away from contextual cues that have already adequately generated the maintenance responses. This diversion of attention generates an error, i.e., a smaller magnitude of activation of the correct maintenance responses. To reduce this error, attention shifts away from the novel cue to the already-known contextual cues. This shift is learned and causes attenuation of learning about the preexposed cue in subsequent phases.

This reconceptualization of the preexposure effect is implemented in the connectionist models in essentially the same way as attenuation of learning after blocking. The implementation includes an input node to represent context cues and an output node to represent maintenance responses. Prior learning establishes a positive associative weight between the contextual cue node and the maintenance response node. The preexposure effect then occurs in the models in the same way as does attenuation after blocking.

This theory of the preexposure effect is admittedly nascent and in need of further support and development far beyond the intended scope of this article. This inchoate theory is offered here for three reasons. First, because Mackintosh's (1975) explanation of the preexposure effect was jettisoned by the connectionist generalizations, a replacement theory was called for. Second, the theory unifies explanations of the preexposure effect with explanations of blocking and other attentional phenomena described below, along with explanations of associative learning. Unlike previous theories, there is a single objective—error reduction—that drives all changes in attention and associative weights. Third, the theory also accounts for some other characteristics of the preexposure effect, and the theory makes predictions that there should be systematic similarities between influences on blocking and influences on the preexposure effect, some of which are mentioned next.

Lubow and Gewirtz (1995) provided a summary of research on the preexposure effect in humans. They noted that for normal adults the preexposure effect is found most robustly when there is some "masking task" that the learner must perform while the preexposed cue is presented. If the masking task is too simple (e.g., if it is not there at all) or if it is too difficult, the preexposure effect will be small or absent. The present theory treats the masking task as a background task much like the ongoing need for maintenance responses elicited by ongoing contextual cues. The difficulty of the masking task corresponds to the amount of attention required for the contextual cues to fully elicit the maintenance responses. If the masking task is simple or automatic, then the contextual cues need only minimal attention to generate a full maintenance response. When another cue is added to the stimulus, attention is not shifted away from the novel cue because the background task does not need attention. If the masking task is very difficult, then the contextual (masking-task) cues garner almost complete attention. When another cue is included in the stimulus, if the cue is of modest salience then attention is not shifted away from the novel cue because it was not able to grab any attention in the first place. It is only when the added cue competes with the masking task for attention that any shift of attention occurs.

The fact that normal adult humans require a masking task (of moderate difficulty) to exhibit the preexposure effect, but animals do not, is accounted for in the present approach by hypothesizing that animals require some amount of attention to carry out background maintenance responses during typical preexposure procedures, but humans can either defer or automatize background maintenance responses during preexposure episodes.

Lubow and Gewirtz (1995) (see also Zalstein-Orda & Lubow, 1995) also summarized research showing that the preexposure and subsequent test (acquisition) phases of the experiment must have the same context in order for the attenuated learning to occur. This *context specificity* is accommodated in the present approach by the exemplar-mediated mapping of cues to attentional gains. The exemplar nodes encode tonic context cues along with phasic cues, so that a learned attentional redistribution is elicited less strongly by a different context.

In so far as the present approach explains attenuation after blocking and the preexposure effect in the same way, the approach predicts that other factors should have analogous influences on both effects. Thus, to the extent that the attention-grabbing ability of the masking task influences the preexposure effect, the relative salience of the cues should influence attenuation after blocking. There has been only a little research on the effects of cue salience in blocking (e.g., Kelin, Weston, McGee-Davis, & Cohen, 1984) and no research on the effects of cue salience on attenuation after blocking. Moreover, to the extent that there is context specificity in the preexposure effect, there should be analogous context specificity in attenuation after blocking. There has been only a small amount of research on the effects of context change in blocking (e.g., Bonardi, Honey, & Hall, 1999, Experiment 3) and no research on the effects of context change on attenuation after blocking.

There is some evidence that blocking and the preexposure effect are not analogously influenced by other factors. For example, it has been reported that blocking is not disrupted by amphetamine (Gray, Pickering, Gray, Jones, Abraham, & Hemsley,

1997), but the preexposure effect is (Gray, Pickering, Hemsley, Dawling, & Gray, 1992). Hippocampal lesions also have different effects on blocking and the preexposure effect (in rats; see Gallo & Candido, 1995). As another example, it has been reported that blocking is weaker in young children than in adults (e.g., Lyczak & Tighe, 1975), but the preexposure effect is exhibited by children even without a masking task (e.g., Kaniel & Lubow, 1986). These apparent discrepancies have not been extensively and systematically explored, however, and a clear view of the commonalities between blocking and the preexposure effect will only be obtained after much future research. Meanwhile, the present theory might prove to be a useful motivation for a number of comparative studies of blocking and the preexposure effect.

There are probably multiple mechanisms underlying blocking and the preexposure effect, and so there might prove to be an attentional mechanism involved in both effects as described by the present theory, along with other mechanisms that differ between the effects. Therefore the attenuated learning after preexposure and after blocking might be affected differently by similar manipulations, despite a common attentional mechanism. Even within the EXIT model, blocking has multiple causes; e.g., learning of summed associative weights and shifts of attention. Other models incorporate representational changes instead of, or in addition to, attentional changes. For example, Saksida (1999) used the attentional learning method of Pearce and Hall (1980) to account for the preexposure effect, along with a competitive learning method for developing enhanced discrimination of preexposed cues to account for perceptual learning (e.g., Honey, Bateson, & Horn, 1994; Gibson & Walk, 1956). Gluck and Myers (1993) explained the preexposure effect without multiplicative attention as used in EXIT or the mixture of experts, instead hypothesizing that preexposure results in a compressed representation of the cue and its context that attenuates the subsequent discrimination of the cue and the context. It would be straightforward to combine the attentional mechanisms of EXIT with the redundancy compression mechanism suggested by Gluck and Myers (1993).

Only future research will determine whether the similarities between attenuated learning after preexposure and after blocking can be attributed to a common attentional mechanism, while the differences between them can be attributed to other underlying mechanisms. Of course, even if a common attentional mechanism exists, the presently offered formalization of it might be inadequate, not only in its details but in its motivating principles.

Interim Summary and Preview

Two connectionist implementations of attentional shifting and learning were fit to data that showed attenuated learning after blocking. Both models use error reduction to drive changes in attention and in associative weights. The effect of the attentional shifting is to accelerate new learning while protecting previous learning.

The models were then shown to have special cases nearly equivalent to Mackintosh's (1975) animal learning model. The connectionist models cannot account for the preexposure effect in the same way as Mackintosh's approach, but

a new theory was offered whereby attenuated learning after preexposure is conceived of as a special case of attenuated learning after blocking.

In the next and final section, I review some of the previous applications of this approach to other phenomena. These applications highlight various behavioral consequences of error-driven attentional shifting and learning.

APPLICATIONS OF THE APPROACH TO OTHER PHENOMENA

In this section I summarize previous applications of connectionist attentional learning models. The goal of this section is to illustrate a range of phenomena (a) in which attentional shifting and learning are implicated and (b) that have been addressed by variants of the connectionist models described above.

Attentional Shifts Facilitate Learning

When people learn a category distinction, they find the task easier, in the sense of achieving greater accuracy faster, when there are fewer relevant dimensions. This fact is most naturally explained by the hypothesis that people learn which dimensions are relevant, and then people shift attention away from irrelevant dimensions. This shift of attention toward relevant dimensions helps differentiate exemplars from different categories, and the shift of attention away from irrelevant dimensions increases similarity within categories. The advantage of fewer relevant dimensions has been reported many times in the literature, but perhaps the simplest demonstration that has been addressed by a connectionist learning model is shown in the left side of Fig. 5, which illustrates two category structures defined on stimuli that vary in two dimensions. The filtration structure has just one relevant dimension, and the other dimension can be ignored without any loss in categorization accuracy. The condensation structure has two relevant dimensions. The right-hand side of Fig. 5 shows data from people who learned the structures, where it can be seen that filtration is much easier than condensation (Kruschke, 1993). These results were fit by the ALCOVE model, which is a connectionist model that learns to shift attention across dimensions by gradient descent on error. Analogous results come from Shepard et al. (1961), who showed that a nonlinearly separable category structure with just two relevant dimensions (their "Type II") is easier to learn than a linearly separable category structure with three relevant dimensions (their "Type IV"). The results were modeled qualitatively using the ALCOVE model by Kruschke (1992), and the results were replicated and modeled quantitatively using the ALCOVE model by Nosofsky, Gluck, Palmeri, McKinley, and Glauthier (1994). When the attentional learning mechanism was shut off by fixing the attentional learning rate to zero, the model could neither show filtration advantage nor the proper ordering of the structures used by Shepard et al. (1961).³

³ When the dimensions cannot be selectively attended to, i.e., when the dimensions are psychologically *integral* as opposed to *separable*, then the advantage of fewer relevant dimensions does not obtain. This is consistent with the theory of learned selective attention. Nosofsky and Palmeri (1996) showed that for integral dimensions a Type-II category structure is more difficult than a Type-IV category structure, and the researchers successfully modeled this result using ALCOVE with its attentional learning rate set to zero.

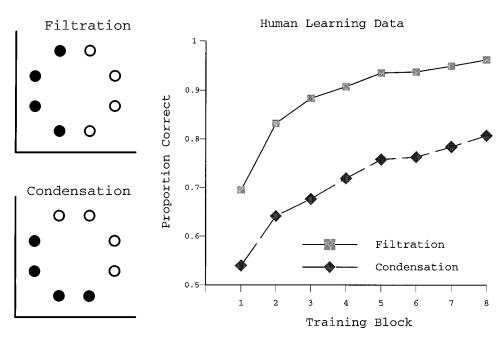


FIG. 5. Filtration and condensation category structures are shown on the left. The axes indicate the two dimensions of psychological space, and the circles indicate particular stimuli within that space. The color of the circle, blank vs filled, indicates which of the two categories was the correct category. On the right is human learning data for the two structures. (Figure adapted from Kruschke, 1993.)

Attentional Shifts Perseverate into Subsequent Learning

Learned attentional shifts perseverate into subsequent phases of learning. When subsequent phases of learning involve the same relevant dimensions, the shift between phases is called an *intradimensional* shift of relevance. When other dimensions are relevant in the subsequent phase, the shift is called *extradimensional*. Many different experiments have shown that intradimensional shifts are easier to learn than extradimensional shifts. This intradimensional advantage is naturally explained in terms of attentional perseveration: When the later phase involves a dimension that has previously been learned to be irrelevant, the learner will have to overcome this suppression of the dimension before it can be learned about.

Traditionally, intradimensional shifts are confounded with changes in novelty. This is because traditional designs have used a single relevant, binary-valued dimension in the initial phase. For example, the initial training might have color being relevant, with red indicating one category and green indicating the other. Creating a new category structure on the same dimension requires introducing new values on the dimension, e.g., blue and yellow. One solution to this problem is to use all novel values on all dimensions (Slamecka, 1968). But this is not optimal because it is difficult to know whether the change in novelty on one dimension is of the same magnitude as the change in novelty on another dimension.

The left panel of Fig. 6 shows a different solution to this problem, such that no novel values are introduced. People initially learned about stimuli that varied on

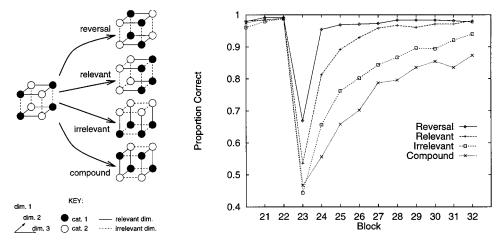


FIG. 6. The left-hand side shows four types of category shifts. The shift type labeled "relevant" is an *intradimensional* shift, and the type labeled "irrelevant" is an *extradimensional* shift. The right side shows the proportion of correct responses during training of each type of shift. The shift occurred at Block 23. (Data from Kruschke, 1996b.)

three binary-valued dimensions, such that two dimensions were relevant. (The category structure is sometimes referred to as an exclusive-OR. The structure is also the same as Type II from Shepard et al., 1961.) There were four possible category structures in the second phase of training, but two are central for the present discussion. The so-called "relevant" shift had just one relevant dimension in the second phase, and this dimension was previously relevant in the first phase. Hence the "relevant" shift is an intradimensional shift. The so-called "irrelevant" shift also had just one relevant dimension in the second phase, but this dimension was irrelevant in the first phase. Hence the "irrelevant" shift is an extradimensional shift. In other aspects the two types of shifts are the same: Both shifts change the categorization of four exemplars, and neither shift introduces any novel dimension values. The right side of Fig. 6 shows data from human learning, where it can be seen that the intradimensional ("relevant") shift is much easier to learn than the extradimensional ("irrelevant") shift. A connectionist model that incorporates attentional learning nicely accommodates the intradimensional shift advantage (Kruschke, 1996b). When the attentional learning mechanism was turned off by fixing the attentional learning rate at zero, the model could not exhibit any intradimensional shift advantage. The only shift type that cannot be addressed by attentional learning alone is the reversal shift. Kruschke (1996b) added a category-label remapping mechanism to address the remarkable ease of reversal shift learning.

Rapid Attentional Shifts Protect Previous Learning

The models described in the present article, namely EXIT and the mixture of experts, emphasized the ability to rapidly shift attention in response to error, before associative weight learning takes place. Rapid shifts of attention are especially important for explaining the *inverse base rate effect*, discovered by Medin and Edelson (1988). In a phased-learning analogue to this effect, people first learn that

symptoms A and B indicate disease 1 (denoted $AB \rightarrow 1$). Later, they continue training with $AB \rightarrow 1$, but also learn cases of $AC \rightarrow 2$. Thus, each disease has one perfectly predictive feature (B predicts 1 and C predicts 2), and the diseases share a symptom, A. In a subsequent test phase, people are presented with novel symptom combinations and asked to make their best diagnoses based on what they learned before. When presented with the shared symptom A by itself, people do not choose diseases 1 and 2 equally, but instead strongly prefer disease 1. When presented with the conflicting symptoms B and C together, again people do not choose the diseases equally, but in this case prefer disease 2 (Kruschke, 1996a, Experiment 2).

This perplexing effect is explained as the consequence of rapidly shifting attention in order to reduce error. Intuitively, the explanation works as follows: In the first phase of learning, people learn that symptoms A and B are associated with disease 1. The strengths of association are moderate, because the two symptoms share predictive roles. In the second phase of learning, people must produce a diagnosis of disease 2 in response to symptoms A and C. If people pay attention to symptom A, they are misled and prone to the wrong diagnosis. To avoid this error, attention shifts to symptom C. Because symptom C is then the main indicator of disease 2, it develops a strong associative weight. Subsequently, when tested with symptom A, the moderate associative weight to disease 1 dominates any weak association with disease 2. When tested with symptoms B and C, the strong association from C to 2 dominates the moderate association from B to 1. A connectionist model (ADIT), which implements rapidly shifting attention, fits quantitative details of the inverse base rate effect (Kruschke, 1996a). The shift of attention facilitates learning of disease 2 while protecting the previously learned association to disease 1.

Recent work has shown that the shift of attention in the inverse base rate effect perseverates into subsequent learning. The idea is that for the symptom pair AC people have learned to attend to symptom C and to pay less attention to symptom A, so that subsequently it should be relatively easy to learn about symptom C and relatively difficult to learn about symptom A. On the other hand, for the symptom pair AB, people have not learned to shift attention away from A so strongly, and so subsequent learning about B will not be much easier than learning about A, or at least the difference between ease of learning about B and A will not be as great as the difference between ease of learning about C and A. These predictions have been verified by recent work (Kruschke, in preparation).

Attentional Shifts Account for "Irrational" Cue Utilization

Rapid shifts of attention, and the learning of these shifts, have also been implicated in a variety of cue-competition phenomena. Consider a situation is which one cue is partially predictive of an outcome, with a 70% probability of the outcome given the cue. The extent to which this cue is utilized turns out to depend on the validity of other cues. If another cue is more valid, then the first cue is utilized less often. If another cue is less valid, then the first cue is utilized more often. This type of cue competition and a variety of related phenomena were successfully accounted for by a model that incorporates: (a) rapid shifts of attention, (b) learning of these shifts, and (c) progressive discounting of probabilistic error

(Kruschke & Johansen, 1999). The rapid shifts of attention and the learning of the shifts were implemented much like they are in the models presented here. The rapid shifts of attention, responding to the probabilistic error, cause the model in many cases to predominantly attend to just one cue or the other. The attention gets stuck on that cue because of the progressive discounting of probabilistic error. The cue with greater validity has a greater probability of attracting attention, at the expense of attention to the other cue.

Attention Can Shift between Internal Representations

In learning classifications of stimuli, people appear to be able to use both exemplars and rules. When classifications are mediated by exemplars, a stimulus is classified according to the graded similarity of the stimulus to known exemplars of the categories. When classifications are mediated by rules, stimuli are classified according to whether or not they satisfy some strict condition, usually involving just a small subset of the possible features. When people learn about a set of stimuli, most of which can be classified according to a rule but some of which are exceptions to the rule, they learn the rule much faster than an exemplar model can account for (Kruschke & Erickson, 1994) and they extrapolate according to the rule instead of according to the nearest exemplar (Erickson & Kruschke, 1998). Nevertheless, the same data also show evidence of exemplar effects. If people use both rules and exemplars, the challenge for modeling is to incorporate both types of representations and a mechanism for allocating use of the different representations as appropriate to different stimuli. One answer to this challenge is a mixture of experts model that has rules in one module and exemplars in another. This sort of model then shifts attention between rules and exemplars in whatever way minimizes error most efficiently. Erickson and Kruschke (1998) and Kruschke and Erickson (1994) reported that this sort of model captures the behavior of human learners much better than an exemplar model alone.4

SUMMARY AND CONCLUSION

Many phenomena in associative learning by both humans and animals are explained by the hypothesis of attentional shifts and the learning of these shifts. This article presented two connectionist formulations of the attentional learning theory. One formulation, called EXIT, modulates the influence of each cue according to an attentional multiplier. The influences of the cues are combined to generate a predicted outcome. The other formulation uses a mixture of experts architecture: Each cue acts as an individual expert, attempting to account for the correct outcomes. In this case, attention modulates the influence of the modules' predictions in the overall prediction of the outcome. In both formulations, the shifting of attention accelerates learning of new associations while protecting previously learned associations.

⁴ Nosofsky and Johansen (2000) show that a modified exemplar model can address some of the results reported by Erickson and Kruschke (1998), but Erickson and Kruschke (2000) demonstrate with new empirical data that this modified exemplar model still cannot handle the basic rule-extrapolation effect.

Both formulations shift attention and learn according to a single principle, i.e., gradient descent on error. There are not separate heuristics for attentional shifting and association learning; instead, there is a unified approach to the two types of change.

Both connectionist formulations were shown to have special cases that are essentially the same as the classic model proposed by Mackintosh (1975). The mixture of experts approach, however, is closer in spirit to Mackintosh's (1975) proposal, in which each cue acts as an independent, unique predictor of the outcome. These formal equivalencies between connectionist models of human learning and a classic model of animal learning represent a step toward a unified model of attention in learning across species.

The models were applied to recent data that show there is learned inattention to a previously blocked cue (Kruschke & Blair, 2000). This attenuation of learning about a previously blocked cue was argued to be analogous to the attenuation of learning about a preexposed cue. The latter effect is referred to as "latent inhibition" or "the preexposure effect." This approach, whereby the preexposure effect is treated as a special case of blocking, is offered here only as a nascent theory, and future research is needed to explore the extent to which it holds true. Nevertheless, the theory is useful as a generator of new research and as a unified approach to two touchstone phenomena that are often treated separately.

A number of other learning phenomena were reviewed which can all be addressed under the unifying umbrella of attentional theory. Rapid shifting of attention and learning of the shifts are good ways to accomplish speedy acquisition of new associations while retaining previously learned associations. This efficiency in learning can lead, however, to seemingly "irrational" behavior in some subsequent situations. The phenomena of blocking and subsequent attenuation of learning, for example, are irrational to the extent that a perfect correlation (between the blocked cue and the outcome) is not learned. The inverse base rate effect, as another example, is irrational to the extent that a truly unpredictive cue is learned to be predictive. Yet these many phenomena are all natural consequences of attentional shifting, which achieves rapid learning of correct responses.

APPENDIX 1: ALTERNATIVE MAPPING OF OUTPUT ACTIVATION TO RESPONSE PROBABILITIES

In situations when the output activations are never negative and at least one is guaranteed to be positive, an alternative mapping of activations to probabilities can be used, instead of the Luce/Softmax rule (Eq. (2)). A version of the GCM proposed by Maddox and Ashby (1993, p. 54; see also Ashby & Maddox, 1993, Proposition 1) raises summed similarity to a power before normalizing. This scheme could be applied to connectionist networks as in the formula for choice probabilities

$$p(c) = \left(a_c^{\text{out}}\right)^{\phi} / \sum_k \left(a_k^{\text{out}}\right)^{\phi},\tag{34}$$

where $\phi > 0$. Nosofsky *et al.* (1994) used this same type of mapping in an extended version of Anderson's (1990) rational model of categorization.

As before (Eq. (2)), a psychological interpretation of the power ϕ in Eq. (34) is that it represents response confidence or decisiveness. When ϕ is large, then relatively small differences in summed similarity produce large response preferences, that is, more decisive responses, with probabilities closer to the extremes of 1 or 0. When ϕ is small, then response probabilities tend to be closer to 0.5. Maddox and Ashby (1993) provided an alternative interpretation in terms of noise in a decision criterion for a deterministic exemplar model. Nosofsky and Palmeri (1997, p. 291) described another interpretation of the power ϕ as a response threshold for a random walk model of speeded classification. Whereas neither of these interpretations might be directly applicable here, they do suggest that this mapping has several possible psychologically meaningful functions, including decisiveness.

The two formalizations of decisiveness (Eqs. (2) and (34)) can be seen to be closely related to each other, after some algebraic manipulation. Note first that Eq. (2) can be reexpressed as.

$$p(c) = \exp(\phi \ a_c^{\text{out}}) / \sum_k \exp(\phi \ a_k^{\text{out}})$$

$$= 1 / \left(1 + \sum_{k \neq c} \exp[-\phi(a_c^{\text{out}} - a_k^{\text{out}})]\right). \tag{35}$$

In the particular case of there being just two categories, c_1 and c_2 , Eq. (35) shows that the choice probability is simply the sigmoidal function of the difference between category activations, $p(c_1) = \text{sig}(a_{c_1}^{\text{out}} - a_{c_2}^{\text{out}})$, where $\text{sig}(x) = 1/[1 + \exp(-\phi x)]$).

Equation (34), on the other hand, can be reexpressed as

$$p(c) = (a_c^{\text{out}})^{\phi} / \sum_k (a_k^{\text{out}})^{\phi}$$

$$= 1 / \left(1 + \sum_{k \neq c} \exp\left[-\phi(\log a_c^{\text{out}} - \log a_k^{\text{out}})\right] \right). \tag{36}$$

In the particular case of there being just two categories, c_1 and c_2 , Eq. (36) shows that the choice probability is simply the sigmoidal function of the difference between the logarithms of the category activations, $p(c_1) = \text{sig}(\log a_{c_1}^{\text{out}} - \log a_{c_2}^{\text{out}})$. Equation (36) is the same as Eqs. (35) except that the output activations are first passed through a logarithmic function (which is possible only if the activations are positive).

The exponential version of decisiveness (Eq. (2)) is used in this article instead of the power version (Eq. (34)) because category activations are possibly negative. To my knowledge, there have not been systematic investigations of the differences between the two formalizations. In at least one application, the two formalizations fit data equally well (Kruschke, Johansen, & Blair, 1999). In any case, it has been shown that some form of flexibility in decisiveness is important for models to fit real data (e.g., Ashby & Gott, 1988; Jones, Wills, & McLaren, 1998; Kruschke *et al.*, 1999).

APPENDIX 2: BASE RATE MIXING IN EXIT IS EQUIVALENT TO BASE RATE MIXING IN ADIT

In the original ADIT model (Kruschke, 1996a), category base rates were mixed with the choice probabilities generated from an associative network to arrive at a base-rate biased, overall choice probability. The mixing of associatively-generated choice probabilities with base-rate choice probabilities was thought to be a distinct mechanism from the Luce/Softmax rule. It is demonstrated now that the mixing rule is actually equivalent to treating the base rates as generated by associative weights from a bias cue, so that no separate mixing mechanism is needed.

The original ADIT model computed output activations and choice probabilities just as in EXIT, using Eqs. (1) $(a_k^{\text{out}} = \sum_i w_{ki} \alpha_i a_i^{\text{in}})$ and (2) $(p(c) = \exp(\phi a_c^{\text{out}})/\sum_k \exp(\phi a_k^{\text{out}}))$. These associatively based choice probabilities were then mixed with the base rates, denoted b(k), according to a multiplicative rule (cf. Kruschke, 1996a, Eq. (11), p. 15) such that the overall probability of choosing category c was given by

$$p'(c) = p(c) b(c)^{\alpha_b} / \sum_{k} p(k) b(k)^{\alpha_b},$$
 (37)

where α_b is, essentially, the attention allocated to the base rates. The attention allocated to the base rates was assumed to decline as the number of stimulus cues increased.

Now I make a new assumption, i.e., that the base rate of a category is learned as an associative weight from a bias cue to the category. The bias cue represents the response prompt and is therefore activated on every trial. It is simple to prove that if a weight learns according to $\Delta w_{ki} = \lambda (t_k - w_{ki}a_i) \, a_i$, then the weight converges to the probability that $t_k = 1$ given $a_i = 1$, assuming that the teacher and input values are either 1 or 0. (This also assumes that the effective learning rate λ does not depend on the input-teacher patterns, which is not necessarily true when attention shifts rapidly to or from the bias cue. Future versions of the model might need to restrict attention shifts on the bias cue.) The output activations produced by these associative weights from the bias cue are then passed through the Luce/Softmax function to produce the choice probabilities generated by the base rates. Thus, the choice probability for category c, generated from the learned base rates, is

$$b(c) = \exp(\phi w_{cb} a_b^{\text{in}}) / \sum_k \exp(\phi w_{kb} a_b^{\text{in}}), \tag{38}$$

where $a_b^{\rm in}=1$ because the bias node, indexed by b, is activated on every trial. Substituting Eq. (38) into Eq. (37) yields, after some simple algebraic manipulation,

$$p'(c) = \frac{\exp(\phi[\sum_{i} w_{ci} \alpha_{i} a_{i}^{\text{in}} + w_{cb} \alpha_{b} a_{b}^{\text{in}}])/}{\sum_{k} \exp(\phi[\sum_{i} w_{ki} \alpha_{i} a_{i}^{\text{in}} + w_{kb} \alpha_{b} a_{b}^{\text{in}}])}.$$
(39)

To reiterate, Eq. (39) reexpresses the formula for mixed choice probabilities from original ADIT in terms of learned associative weights from a bias node.

A bias node can also be implemented in EXIT, like any other cue node. Let the bias node be indexed by subscript b. Then the output activation in EXIT is determined by Eq. (1), which becomes

$$a_k^{\text{out}} = \sum_i w_{ki} \, \alpha_i a_i^{\text{in}} + w_{kb} \, \alpha_b \, a_b^{\text{in}} \,. \tag{40}$$

When this is substituted into EXIT's formula for choice probability (the Luce/Softmax rule, Eq. (2)), the resulting formula is immediately the same as Eq. (39).

In conclusion, the original multiplicative mixing rule in ADIT is essentially the same as learning base rates from a bias cue and treating this cue like any other cue. Thus, the multiplicative base rate mixing rule of Eq. (37) is not a different mechanism from activation summation in the output nodes.

REFERENCES

- Aha, D. W., & Goldstone, R. (1990). Learning attribute relevance in context in instance-based learning algorithms. In *Proceedings of the twelfth annual conference of the cognitive science society* (pp. 141–148). Hillsdale, NJ: Erlbaum.
- Anderson, J. R. (1990). The adaptive character of thought. Hillsdale, NJ: Erlbaum.
- Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14, 33–53.
- Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, 37, 372–400.
- Bonardi, C., Honey, R. C., & Hall, G. (1999). Context specificity of conditioning in flavor-aversion learning: Extinction and blocking tests. Animal Learning and Behavior, 18(3), 229–237.
- Brandon, S. E., & Wagner, A. R. (1998). Occasion setting: Influences of conditioned emotional responses and configural cues. In N. A. Schmajuk & P. C. Holland (Eds.), Occasion setting: Associative learning and cognition in animals (pp. 343–382). Washington, D.C.: American Psychological Association.
- Bridle, J. S. (1990). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In F. Fogelman-Soulié & J. Hérault (Eds.), Neurocomputing: Algorithms, architectures and applications (pp. 227–236). New York: Springer-Verlag.
- Dickinson, A. (1980). Contemporary animal learning theory. Cambridge, UK: Cambridge University Press
- Domjan, M. (1998). The principles of learning and behavior (4th ed). Pacific Grove, CA: Brooks/Cole.
- Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, **127**, 107–140.
- Erickson, M. A., & Kruschke, J. K. 2000. Rule-based extrapolation in perceptual categorization, submitted for publication.
- Estes, W. K. (1988). Toward a framework for combining connectionist and symbol-processing models. *Journal of Memory and Language*, 27, 196–212.
- Estes, W. K. (1994). Classification and cognition. New York: Oxford University Press.
- Frey, P. W., & Sears, R. J. (1978). Model of conditioning incorporating the Rescorla-Wagner associative axiom, a dynamic attention process, and a catastrophe rule. *Psychological Review*, **85**, 321–340.
- Gallo, M., & Candido, A. (1995). Dorsal hippocampal lesions impair blocking but not latent inhibition of taste aversion learning in rats. *Behavioral Neuroscience*, 109(3), 413–425.
- Garner, W. R. (1974). The processing of information and structure. Hillsdale, NJ: Erlbaum.
- Gibson, E. J., & Walk, R. D. (1956). The effect of prolonged exposure to visually presented patterns on learning to discriminate them. *Journal of Comparative and Physiological Psychology*, **49**, 239–242.

- Gluck, M. A., & Bower, G. H. (1988a). Evaluating an adaptive network model of human learning. *Journal of Memory and Language*, 27, 166–195.
- Gluck, M. A., & Bower, G. H. (1988b). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, 117, 227–247.
- Gluck, M. A., & Myers, C. E. (1993). Hippocampal mediation of stimulus representation: a computational theory. *Hippocampus*, 3, 491–516.
- Gray, N. S., Pickering, A. D., Gray, J. A., Jones, S. H., Abrahams, S., & Hemsley, D. R. (1997). Kamin blocking is not disrupted by amphetamine in human subjects. *Journal of Psychopharmacology*, 11(4), 301–311.
- Gray, N. S., Pickering, A. D., Hemsley, D. R., Dawling, S., & Gray, J. A. (1992). Abolition of latent inhibition by a single 5 mg dose of d-amphetamine in man. *Psychopharmacology*, 107, 425–430.
- Honey, R. C., Bateson, P., & Horn, G. (1994). The role of stimulus comparison in perceptual learning: An investigation with the domestic chick. *Learning and Motivation*, **20**, 262–277.
- Jacobs, R. A. (1997). Nature, nurture, and the development of functional specializations: A computational approach. *Psychonomic Bulletin & Review*, **4**, 299–309.
- Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. Neural Computation, 3, 79–87.
- Jones, F. W., Wills, A. J., & McLaren, I. P. L. (1998). Perceptual categorization: Connectionist modeling and decision rules. *Quartely Journal of Experimental Psychology: Comparative and Physiological Psychology*, 51B(1), 33–58.
- Kamin, L. J. (1968). "Attention-like" processes in classical conditioning. In M. R. Jones (Ed.), Miami symposium on the prediction of behavior: Aversive stimulation (pp. 9–33). Coral Gables, FL: University of Miami Press.
- Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment* (pp. 279–296). New York: Appleton-Century-Crofts.
- Kaniel, S., & Lubow, R. E. (1986). Latent inhibition: A developmental study. British Journal of Developmental Psychology, 4, 367–375.
- Klein, S. B., Weston, D., McGee-Davis, T., & Cohen, L. (1984). The relative contributions of predictiveness and salience in flavor aversion learning. *Learning and Motivation*, 15, 188–202.
- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, **99**, 22–44.
- Kruschke, J. K. (1993). Human category learning: Implications for back propagation models. Connection Science, 5, 3–36.
- Kruschke, J. K. (1996a). Base rates in category learning. Journal of Experimental Psychology: Learning, Memory & Cognition, 22, 3–26.
- Kruschke, J. K. (1996b). Dimensional relevance shifts in category learning. Connection Science, 8, 201–223.
- Kruschke, J. K. (in preparation). Learning involves attention. In G. Houghton (Ed.), Connectionist models in cognitive psychology. London: Psychology Press.
- Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. Psychonomic Bulletin & Review, 7, 636–645.
- Kruschke, J. K., & Bradley, A. L. (1995). Extensions to the delta rule for human associative learning. Indiana University Cognitive Science Research Report 141 [available at http://www.indiana.edu/~kruschke/kb95abstract.html].
- Kruschke, J. K., & Erickson, M. A. (1994). Learning of rules that have high-frequency exceptions: New empirical data and a hybrid connectionist model. In *The proceedings of the sixteenth annual conference of the cognitive science society* (pp. 514–519). Hillsdale, NJ: Erlbaum.
- Kruschke, J. K., & Erickson, M. A. (1995). Six principles for models of category learning. Talk presented at the *36th annual meeting of the Psychonomic Society*, *10 November 1995*, *Los Angeles*, *CA* [available via WWW at http://www.indiana.edu/~kruschke/psychonomics95_abstract.html].

- Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory & Cognition*, **25**, 1083–1119.
- Kruschke, J. K., Johansen, M. K., & Blair, N. J. (1999). Exemplar model account of inference learning: Comment on Yamauchi and Markman (1998) [available at http://www.indiana.edu/~kruschke/yamauchicomment.html].
- Lawrence, D. H. (1949). Acquired distinctiveness of cues. I. Transfer between discriminations on the basis of familiarity with the stimulus. *Journal of Experimental Psychology*, 39, 770–784.
- Lawrence, D. H. (1950). Acquired distinctiveness of cues. II. Selective association in a constant stimulus situation. *Journal of Experimental Psychology*, 40, 175–188.
- Lubow, R. E. (1989). Latent inhibition and conditioned attention theory. Cambridge, UK: Cambridge University Press.
- Lubow, R. E., & Gewirtz, J. C. (1995). Latent inhibition in humans: Data, theory, and implications for schizophrenia. *Psychological Bulletin*, 117(1), 87–103.
- Lubow, R. E., & Moore, A. U. (1959). Latent inhibition: The effect of nonreinforced preexposure to the conditioned stimulus. *Journal of Comparative and Physiological Psychology*, 52, 415–419.
- Luce, R. D. (1959). Individual choice behavior. New York: Wiley.
- Lyczak, R., & Tighe, T. (1975). Stimulus control in children under a blocking paradigm. Child Development, 46, 115–122.
- Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, 82, 276–298.
- Mackintosh, N. J., & Turner, C. (1971). Blocking as a function of novelty of CS and predictability of UCS. Quarterly Journal of Experimental Psychology, 23, 359–366.
- Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception and Psychophysics*, 53, 49–70.
- Massaro, D. W. (1987). Speech perception by ear and eye: A paradigm for psychological inquiry. Hillsdale, NJ: Erlbaum.
- McLaren, I. P. L., Kaye, H., & Mackintosh, N. J. (1989). An associative theory of the representation of stimuli: applications to perceptual learning and latent inhibition. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology* (pp. 102–130). Oxford, UK: Oxford University Press.
- Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General*, 117, 68–85.
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. Psychological Review, 85, 207–238.
- Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. Psychological Bulletin, 117, 363-386.
- Moore, J. W., & Stickney, K. J. (1980). Formation of attentional-associative networks in real time: Role of the hippocampus and implications for conditioning. *Physiological Psychology*, **8**, 207–217.
- Nosofsky, R. M. (1986). Attention, similarity and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115, 39–57.
- Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, 22, 352–369.
- Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review*, 7, 375–402.
- Nosofsky, R. M., & Palmeri, T. J. (1996). Learning to classify integral-dimension stimuli. Psychonomic Bulletin & Review, 3, 222–226.
- Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, 104, 266–300.

- Pearce, J. M., & Hall, G. H. (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 non-reinforcement. 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., Durbin, R., Golden, R., & Chauvin, Y. (1995). Backpropagation: The basic theory. In Y. Chauvin & D. E. Rumelhart (Eds.), *Backpropagation: Theory, architectures, and applications* (pp. 1–34). Hillsdale, NJ: Erlbaum.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing* (Vol. 1, pp. 318–362). Cambridge, MA: MIT Press.
- Saksida, L. M. (1999). Effects of similarity and experience on discrimination learning: A nonassociative connectionist model of perceptual learning. *Journal of Experimental Psychology: Animal Behavior Processes*, 25, 308–323.
- Schmajuk, N. A., Lam, Y.-W., & Gray, J. A. (1996). Latent inhibition: A neural network approach. Journal of Experimental Psychology: Animal Behavior Processes, 22(3), 321–349.
- Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, 1, 54–87.
- Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, **237**, 1317–1323.
- Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. Psychological Monographs, 75(13) [Whole No. 517].
- Siegel, S., & Allan, L. G. (1996). The widespread influence of the Rescorla-Wagner model. *Psychonomic Bulletin & Review*, **3**, 314–321.
- Slamecka, N. J. (1968). A methodological analysis of shift paradigms in human discrimination learning. Psychological Bulletin, 69, 423–438.
- Sutherland, N. S., & Mackintosh, N. J. (1971). Mechanisms of animal discrimination learning. New York: Academic Press.
- Trabasso, T., & Bower, G. H. (1968). Attention in learning: Theory and research. New York: Wiley.
- 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.
- Wickens, T. D. (1989). Multiway contingency tables analysis for the social sciences. Hillsdale, NJ: Erlbaum.
- Williams, B. A. (1999). Associative competition in operant conditioning: Blocking the responsereinforcer association. *Psychonomic Bulletin & Review*, 6, 618–623.
- Zalstein-Orda, N., & Lubow, R. E. (1995). Context control of negative transfer induced by preexposure to irrelevant stimuli: Latent inhibition in humans. *Learning and Motivation*, **26**(1), 11–28.

Received: June 7, 1999; published online: June 6, 2001