

# Attention in Learning

John K. Kruschke<sup>1</sup>

Department of Psychology, Indiana University, Bloomington, Indiana

## Abstract

Learners exhibit many apparently irrational behaviors in their use of cues, sometimes learning to ignore relevant cues or to attend to irrelevant ones. A learning phenomenon called highlighting seems especially to demand explanation in terms of learned attention. Highlighting complements the classic phenomenon of conditioned blocking, which has been shown to involve learned inattention. Highlighting and blocking, along with a wide spectrum of other perplexing learning phenomena, can be accounted for by recent connectionist models in which both attentional shifting and associative learning are driven by the rational goal of rapid error reduction.

## Keywords

attention; blocking; highlighting; learning; connectionist model

Will tomorrow bring rain or sunshine? The weather might be predicted by myriad cues: the color of the sunset, the shapes of the clouds, the direction the cows are facing, the number of neighbors washing their cars. Which cues should an observer attend to? People can shift their attention to the cues that predict the weather, and people can learn to associate those cues with specific outcomes such as rain or shine.

Researchers have long studied how humans and animals learn to allocate attention across potentially informative cues (e.g., Trabasso &

Bower, 1968), and the past decade has produced notable empirical and theoretical advances. In this article, I focus on two such advances: an effect called *highlighting* that offers new opportunities for studying attentional learning, and the development of connectionist learning models that explain both attentional shifting and associative learning by using the unifying mechanism of error reduction.

## HIGHLIGHTING: A NEW ATTENTIONAL EFFECT

Three phenomena have often been interpreted as reflecting attentional learning (for a review, see Oades & Sartory, 1997). One is latent inhibition, which occurs when a cue is initially presented with no apparent outcome but later is made to be a perfect predictor of a novel outcome. People are slow to learn about the cue's predictiveness, apparently because they inhibit attention to the cue (Lubow, 1989). A second attentional effect is difficulty with extradimensional relevance shifts.<sup>2</sup> In an extradimensional relevance shift, the predictiveness of two cue dimensions changes. Initially cues from one dimension (e.g., color) are relevant and cues from a second dimension (e.g., shape) are irrelevant, but then the relevance of the dimensions is exchanged. This type of shift is difficult to learn, apparently because people have learned to ignore the initially irrelevant dimension. A third phenomenon of attentional learning is conditioned blocking. I review blocking in detail to provide context for a complementary phenom-

enon, called highlighting, that I propose should be added to the list of attentional effects.

## Blocking

Conditioned blocking refers to a situation in which a new cue accompanies an old cue that was already learned to perfectly predict an outcome. People tend not to associate the new cue with the outcome; that is, learning about the new cue has apparently been blocked. The blocking effect has been produced in a wide variety of paradigms and species since its discovery by Kamin (1968). In one type of blocking procedure, human participants are asked to view lists of symptoms of hypothetical patients, one at a time, and to guess for each patient the appropriate diagnosis from a menu of (fictitious) diseases. After each guess, the correct diagnosis is provided, and the learner is allowed to study the symptoms and correct diagnosis before moving on to the next case. The structure of the pairings between cues (symptoms) and outcomes (diseases), and the phases of training, are shown in Table 1. In the early phase of training, the learner studies cases in which cue A produces outcome X, denoted as  $A \rightarrow X$ . In a later phase of training, the cue is accompanied by a redundant relevant cue, B, but still leads to the same outcome, X (denoted by  $A.B \rightarrow X$ ). The transition from early to late phases of training is seamless.

The later training phase also includes intermixed trials of cases, denoted as  $C.D \rightarrow Y$ , in which two different symptoms lead to a second outcome. These control cases occur with the same frequency as  $A.B \rightarrow X$  trials. If all that matters for the learning of associations is the number of co-occurrences, then the strength of association between B and X should be the same as the

**Table 1.** *Designs for blocking and highlighting procedures*

Phase	Procedure	
	Blocking	Highlighting
Early training	A → X	A.B → X
Late training	A.B → X, C.D → Y	A.B → X, A.D → Y
Testing	B.D → ? (Result: Y is preferred; i.e., B is blocked)	B.D → ? (Result: Y is preferred; i.e., D is highlighted)

*Note.* A, B, C, and D are stimulus cues. X and Y are outcomes. Arrows indicate the correspondence of cues to outcomes; question marks indicate trials on which the correct response is not provided.

strength of association between D and Y.

This prediction of equal associative strength is assessed in the final testing phase, in which cues B and D are presented together and the learners are asked to make their best diagnosis based on what they learned previously. The result is a strong tendency to choose outcome Y rather than outcome X. Apparently, learning about cue B has been prevented, or "blocked."

The discovery of blocking, and its ubiquity, had an enormous impact on theories of learning. It disconfirmed simple contiguity theories, according to which learning is determined by the frequency of co-occurrences of cue and outcome, and it posed a major challenge for new theories.

The dominant interpretation of blocking posits that the outcome X has a diminished influence on learning when the participants study cases of A.B → X in the later phase of training. This theory, which was formalized in the enormously influential mathematical model developed by Rescorla and Wagner (1972), suggests that because the learner has already learned that cue A fully predicts outcome X, the learner experiences little surprise when cases of A.B → X appear, and hence learns little about cue B.

Blair and I (Kruschke & Blair, 2000) provided evidence that this explanation is incomplete because people do, in fact, learn something about the redundant relevant cue: They learn to ignore it. We showed that subsequent learning of a new association with the blocked cue is retarded compared with learning of a new association with an unblocked control cue. The retardation of learning is predicted by the conventional assumption that a learner will take a longer time to learn about an attentionally suppressed cue than about a cue that is attended. Whereas blocking per se can be explained without appeal to attention, the subsequent retardation of learning about a blocked cue cannot be accounted for by the Rescorla-Wagner model and is most naturally explained as learned inattention.

This new empirical result is the first such demonstration in humans and is generally consistent with the attentional theories of Mackintosh (1975), whose particular formalism has been generalized and unified within a new connectionist framework described later in this article.

### Highlighting

Whereas in blocking people learn to ignore a newly added cue,

in highlighting people learn to attend to it. The right side of Table 1 shows the basic structure of training that produces the highlighting effect (adapted from the "inverse base-rate effect" reported by Medin & Edelson, 1988). In the early phase of training, cues A and B are presented together and indicate outcome X. In later training, cases of A.B → X continue, but are intermixed with cases of A.D → Y. The structure of the two cases is symmetrical in that each outcome has a single perfect predictor (B for X and D for Y) and the outcomes share an imperfect predictor (A).

If people learn the structural symmetry, then cue A should not be differentially associated with the outcomes, and cues B and D should be equally associated with their respective outcomes, X and Y. In particular, the Rescorla-Wagner (1972) model predicts that learning should be symmetric after adequate learning in the later phase.

These predictions are disconfirmed by results from the final testing phase. When cue A is presented by itself, people strongly prefer outcome X. When cues B and D are presented together, people reliably prefer outcome Y. Apparently, when learning case A.D → Y, people shift attention away from A, which they have already learned indicates the other outcome X, and attentionally highlight cue D. The phenomenon of highlighting occurs across a number of variations of the training procedures and stimuli (see citations in Kruschke, 2001a).

The phenomenon of highlighting implicates rapid shifts of attention during encoding; that is, on a given learning trial, the attentional shift is made largely before the associations are learned. On a later-trained trial of A.D → Y, attention rapidly shifts away from A toward D to achieve the preservation of the previously learned association of A with X, and the prevention of a du-

plicitous association of A with Y. This rapid attentional shift greatly reduces interference between the previously learned case of  $A.B \rightarrow X$  and the to-be-learned case of  $A.D \rightarrow Y$ . If the shift of attention were not rapid, then the initial association of A with X would be extinguished and the association of D with Y would not differ much in strength from the association of B with X.

The attentional shift must itself be learned, however, so that when case A.D appears, attention is shifted away from A to D. This learned attention can be assessed by examining learning subsequent to highlighting. It is conventionally assumed that learning about an ignored cue is retarded. Thus, after highlighting, subsequent learning about cue A should be slower than learning about cue D, specifically when cues A and D are presented together. Unpublished experiments in my lab have confirmed this prediction.

### ATTENTIONAL SHIFTING BY ERROR REDUCTION

Blocking and highlighting are just two examples of seemingly irrational learning. In both blocking and highlighting procedures, cues B and D are equally perfect predictors of their respective outcomes (i.e.,  $p(X|B) = p(Y|D) = 100\%$ ). Nevertheless, in blocking, B is apparently underweighted, and in highlighting, D is apparently overweighted. In both procedures, there is evidence that people have shifted attention to specific cues and learned those attentional reallocations.

It turns out that these attentional shifts, and their concomitant irrational behaviors, can be accurately modeled by a simple rational process: rapid error reduction. Consider first the case of highlighting.

When learning  $A.B \rightarrow X$  in the early phase of training, people attend to both cues equally (on average) by default. In the later training phase, when people are learning  $A.D \rightarrow Y$ , any attention to A generates an erroneous production of response X. They reduce this error by shifting attention away from A to D.

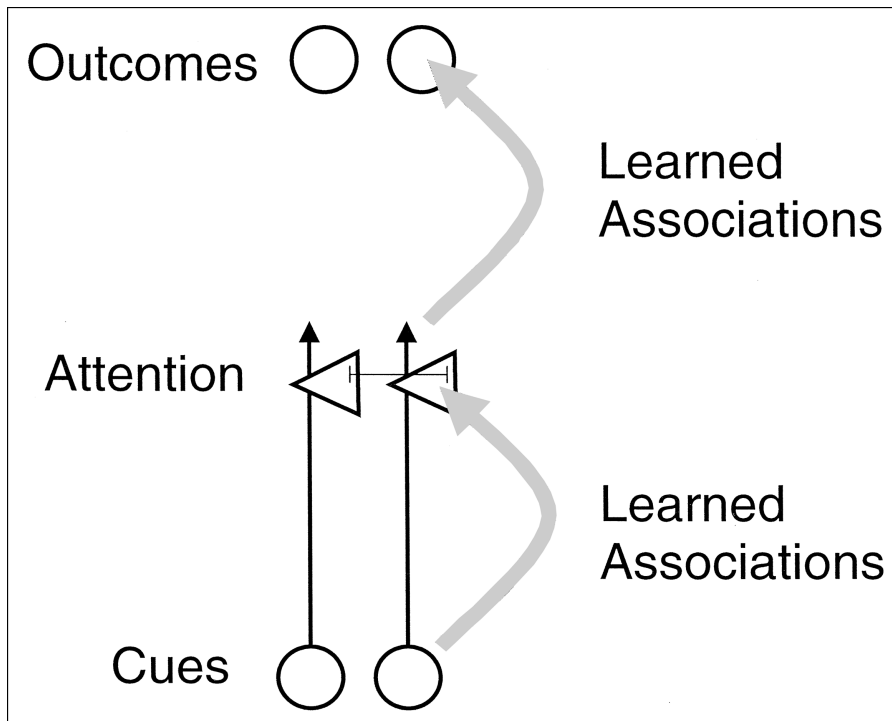
Consider next the case of blocking. When the first cases of  $A.B \rightarrow X$  appear in the late-training phase, B garners some attention by default. Because attention has limited capacity, this distraction by B causes A to get less than full attention, and, therefore, the previously learned X response is not generated as strongly as it should be. This error can be quickly corrected if attention is shifted away from B to A.

This general scheme of attentional shifting by error reduction has been rigorously implemented in a series of specific connectionist models developed in recent years (e.g., Kruschke, 2001b; Kruschke & Johansen, 1999). In connectionist models generally, cues and outcomes are represented by activations of nodes in a network, roughly analogous to interconnected neurons in the brain. Activation flows from node to node via connections with different weights, roughly analogous to neural synapses with different conductances. The connectionist models of attentional shifting generalize and unify historically influential models, such as the Rescorla-Wagner (1972) model, the attentional model of Mackintosh (1975), and the generalized context model of Nosofsky (1986).

Figure 1 is a schematic of the architecture common to all these connectionist networks. When cues are presented, they activate nodes at the lowest layer of the network. Attention is represented by multiplicative gates (denoted in the figure by triangles) on the cue activations.

The input activations are transmitted (along the thin arrows in the figure) to these gates, where they are made available for attentional modulation; attended-to cues are amplified by large multipliers (i.e., gates with high activations) and ignored cues are attenuated by small multipliers (i.e., gates with low activations). Any presented cue garners some attention by default, and the attentional gates compete (denoted in Fig. 1 by the horizontal line between triangles) for a limited-capacity pool of activation. The attentionally gated cue activation then propagates to outcome nodes at the top of the network. If a cue has a small attentional multiplier, its activation is not propagated strongly to the outcome nodes. The activations at the outcome nodes are transformed into choice probabilities to match human choice data.

Learning in this architecture proceeds as follows: When cues are presented at the beginning of a trial, the corresponding cue nodes are given activation values of 1, whereas nodes for absent cues have zero activation. These activations are transmitted to the attentional gates via the fixed connections represented by the thin arrows in Figure 1, and at the same time learned connections from the cues to the attentional gates (the lower thick arrow in Fig. 1) differentially activate the gates, thereby allocating learned attention across the cues. These learned connections thus generate the distribution of attention. Activation then propagates from the attentionally gated cues to the outcome nodes (via the upper thick arrow in Fig. 1). The activation of the outcome nodes is transformed into a choice, such that more highly activated outcomes are more likely to be chosen than less activated outcomes. This choice corresponds to a learner's response in an environment. Then the correct response is presented to



**Fig. 1.** Connectionist architecture for learning attention and outcomes. Error at the outcome nodes initially drives rapid shifting of attention (at the triangle-shaped nodes). After attention is shifted, the remaining error drives learning of associations (thick arrows) between the cues and the shifted attentional distribution and between the attentionally gated cues and the outcomes. The associative links can be direct or mediated by intervening layers of nodes.

the network, just as corrective feedback is presented to learners in the experiments. The correct response is represented in the network by desired activation values of the outcome nodes; the correct outcome has a desired activation value of 1 and the other, incorrect outcomes have a desired activation value of 0. The model computes the discrepancy between its guessed outcome activations and the correct activations. The singular goal of the model is to reduce this error.

The first step in reducing the error is a rapid shift of attention away from cues that cause error and toward cues that reduce error. This attentional shift corresponds to a change in the activations of the attentional multipliers so that the desired distribution of attention across the current cues will be achieved. The model should learn

to generate this desired attentional distribution, instead of the distribution it erroneously generated, whenever these cues are presented. The model computes the discrepancy between the desired attentional distribution and the current attentional distribution, and then adjusts the associative weights from the cues to the attention nodes to approximate the desired distribution (the lower thick arrow in Fig. 1). Finally, the associative weights to the outcome nodes (upper thick arrow in Fig. 1) are adjusted to reduce the remaining error. Thus, both attention shifting and association learning are driven by the goal of error reduction.

Different instantiations of this general architecture have been used when addressing different subsets of learning phenomena. The RASHNL model (Kruschke &

Johansen, 1999) placed a layer of nodes between the attentional gates and the outcome nodes to mediate complex mappings from cues to outcomes. The EXIT model (Kruschke, 2001b) employed a layer of nodes between the input cues and the attentional gates. A goal for future research is to combine these architectural features and address all the phenomena simultaneously.

The RASHNL and EXIT models and related models have been shown to accurately fit data from a wide variety of experiments, including not only the blocking and highlighting effects, but also many other seemingly irrational phenomena in learning, such as under- or overutilization of partially predictive cues, relevance and reversal shifts, and differential difficulty of learning different category structures. The models have also made novel predictions (e.g., they predicted the deleterious impact of undiagnostic but salient cues; Kruschke & Johansen, 1999, Experiments 3 and 4) and have unified a number of otherwise distinct effects (e.g., latent inhibition and blocking; Kruschke, 2001b). Thus, a spectrum of apparently irrational learning phenomena can be accounted for, in rigorous detail, by the rational goal of rapid error reduction.

## RAMIFICATIONS

Phenomena and theories of attentional shifting promise to have an impact on many fields. Highlighting may play a role in stereotype formation, which is studied by social psychologists. Highlighting and retarded learning after blocking may be used to assess dysfunctional attention, a concern of clinical psychologists. When the models are fit to learning data from people with disorders such as

schizophrenia, Huntington's disease, or Parkinson's disease, the models' attentional-shifting rate and associative-learning rates may help identify the aspects of learning that differ between these and normal populations. Similarly, when the models are fit to learning data from different age groups, the models' attentional-shifting rate and associative-learning rates may help identify which aspects of learning change through childhood, adolescence, and maturity. Consumers and marketers will be interested in how attentional shifting might influence assessments of products and consumer choice. Educators and trainers will want to know how best to arrange topics to maximize efficiency in learning while minimizing biases such as blocking and highlighting. In summary, the study of attention in associative learning holds promise for unifying disparate, perplexing behavioral effects, explaining those effects in detail through formal models, and applying the findings to many other fields.

### Recommended Reading

Kruschke, J.K. (2001b). (See References)  
Mackintosh, N.J. (1975). (See References)

Oades, R.D., & Sartory, G. (1997). (See References)

**Acknowledgments**—The author's research has been supported by National Institute of Mental Health FIRST Award R29-MH51572 and National Science Foundation Grant BCS-9910720.

### Notes

1. Address correspondence to John K. Kruschke, Department of Psychology, Indiana University, 1101 E. 10th St., Bloomington, IN 47405-7007; e-mail: kruschke@indiana.edu.

2. Recent evidence from behavioral effects, modeling, and neuroscience has shown that relevance shifts use different mechanisms than reversal shifts, in which the same cues remain relevant but the correct outcomes are reversed (see Kruschke, 1996; Rogers, Andrews, Grasby, Brooks, & Robbins, 2000). Other work has begun to separate effects of learned relevance from learned irrelevance (see Owen et al., 1993).

### References

- 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.
- Kruschke, J.K. (1996). Dimensional relevance shifts in category learning. *Connection Science*, 8, 201-223.
- Kruschke, J.K. (2001a). The inverse base rate effect is not explained by eliminative inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 1385-1400.
- Kruschke, J.K. (2001b). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology*, 45, 812-863.
- Kruschke, J.K., & Blair, N.J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review*, 7, 636-645.
- Kruschke, J.K., & Johansen, M.K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 1083-1119.
- Lubow, R.E. (1989). *Latent inhibition and conditioned attention theory*. Cambridge, England: Cambridge University Press.
- Mackintosh, N.J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, 82, 276-298.
- 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.
- Nosofsky, R.M. (1986). Attention, similarity and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115, 39-57.
- Oades, R.D., & Sartory, G. (1997). The problems of inattention: Methods and interpretations. *Behavioural Brain Research*, 88, 3-10.
- Owen, A.M., Roberts, A.C., Hodges, J.R., Summers, B.A., Polkey, C.E., & Robbins, T.W. (1993). Contrasting mechanisms of impaired attentional set-shifting in patients with frontal lobe damage or Parkinson's disease. *Brain*, 116, 1159-1175.
- 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.
- Rogers, R.D., Andrews, T.C., Grasby, P.M., Brooks, D.J., & Robbins, T.W. (2000). Contrasting cortical and subcortical activations produced by attentional-set shifting and reversal learning in humans. *Journal of Cognitive Neuroscience*, 12, 142-162.
- Trabasso, T., & Bower, G.H. (1968). *Attention in learning: Theory and research*. New York: Wiley.