Can Causal Induction Be Reduced to Associative Learning?

Michael R. Waldmann
University of Frankfurt

Keith J. Holyoak
University of California, Los Angeles

ABSTRACT

A number of researchers have recently claimed that higher-order human learning, such as categorization and causal induction, can be explained by the same principles as govern lower-order learning, such as classical conditioning in animals. An alternative view is that people often impose abstract causal models on observations, rather than simply associating inputs with outputs. We report three experiments using a multiple-cue learning paradigm in which models based on associative learning versus abstract causal models make opposing predictions. We show that different causal models can yield radically different learning from identical observations. In particular, we compared people's abilities to learn when the positive cases were defined by a linear cue-combination rule versus a rule involving a within-category correlation between cues. The linear structure was more readily learned when the cues were interpreted as possible causes of an effect to be predicted, whereas the correlated structure was more readily learned when the cues were interpreted as the effects of a cause to be diagnosed. The results disconfirm all associative models of causal induction in which inputs are associated with outputs without regard for causal directionality.

Introduction

The Associative View of Multiple-Cue Learning

Tasks as different as classification learning, causal induction, and classical conditioning can be viewed as examples of multiple-cue learning. In each of these tasks, a number of cues, which might be features, causes, or conditional stimuli, are combined to trigger a response. This response might be a classification decision, a prediction of an effect, or a conditioned response. Because of the apparent similarity between different types of multiple-cue learning situations, it is tempting to postulate common underlying learning mechanisms for them. A currently popular view of multiple-cue learning treats it as a bottom-up process that is fundamentally associative in nature. Thus higher-order types of learning in humans, such as classification learning, and lower-order types of learning in animals, such as classical conditioning, are seen as examples of similar learning processes.

A number of researchers have recently claimed that higher-order types of human learning, such as categorization and causal induction, can be explained by principles that govern lower-order learning in animals, such as classical conditioning (e.g., Gluck & Bower, 1988a, b; Shanks & Dickinson, 1987). In particular, Gluck and Bower (1988a, b) have suggested that adaptive associative networks can provide powerful models of human categorization as well as of classical conditioning. These connectionist models consist of an input layer that represents potential cues, such as symptoms of possible diseases observed in a patient, and an output layer that might represent classification responses, such as diagnoses of alternative diseases. The responses are computed by a linear function of the weighted cues. The weights are learned in a competitive fashion using the least mean squares (LMS) learning rule (Widrow & Hoff, 1960), in which the weights are incrementally updated in proportion to the response error they produce. Gluck and Bower have shown that a simple model of this sort compares favorably with other models of human categorization (also see Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; for a critique see Shanks, 1990). Since the LMS rule is formally equivalent to

This research was supported by NSF Grant BNS 87-10305 to Patricia Cheng. Michael Waldmann was sponsored by a grant from the German Research Foundation. We thank Patricia Cheng for helpful advice and discussions.

Rescorla and Wagner's (1972) theory of classical conditioning (Sutton & Barto, 1981), these findings suggest that important commonalities link higher-order learning such as categorization and lower-order classical conditioning. Thus Gluck and Bower basically claim that categorization can be modelled as simple associative learning. Similarly, Shanks and Dickinson (1987) argue that causal induction can be reduced to associative learning.

As pointed out by Minsky and Papert (1969), simple one-layer networks can only learn linearly-separable learning tasks. To deal with this major limitation, various extensions of associative network models have been suggested in the connectionist literature. Gluck, Hee, and Bower (1989) proposed a configural-cue network in which pairwise conjunctions of simple cues are coded using configural cues added to the input layer (see also Gluck & Bower, 1988b). Alternatively, the standard connectionist approach to nonlinear learning tasks is to add intermediate layers of hidden nodes between the input and the output layers, which can be used to code cue combinations (Rumelhart, Hinton, & Williams, 1986). Using backpropagation of error signals, which conceptually is an extension of the LMS learning rule to multiple-layer networks, these hidden units can be trained to code interactions in the input. Despite the differences among the various alternative network models, each of these connectionist learning schemes shares a fundamental associationistic assumption: The network simply tries to learn statistical associations between the nodes coded on the input level and the desired output.

Learning Within Abstract Causal Models

The associative view of learning can be contrasted with a more mentalistic approach, which can be traced back to Gestalt psychology. In this tradition, it is claimed that people use abstract, meaningful world knowledge to guide their learning about new domains. Higher-order learning and lower-order associative learning are seen as different in important ways. In particular, one view of human learning is that people impose abstract causal models on observations. Wattenmaker, Dewey, Murphy, and Medin (1986) have shown that people profit from specific world knowledge. People become more sensitive to structural relations between the input cues during learning when they can relate the learning material to previously acquired knowledge. We will argue here that even in situations in which people cannot bring to bear specific world knowledge, they nonetheless might use abstract knowledge about central properties of the world -- in particular, abstract knowledge about causal relations. We have set up an experimental situation in which associative learning and learning based on abstract causal models can be pitted against each other. We will show that different causal models can yield radically different learning from identical observations, a finding that cannot be explained by associative learning models.

Figure 1 illustrates how we decouple higher-order causal learning from associative learning in our experiments. The arrows represent temporal precedence, either in order of presentation of the information, or in order of cause and effect.

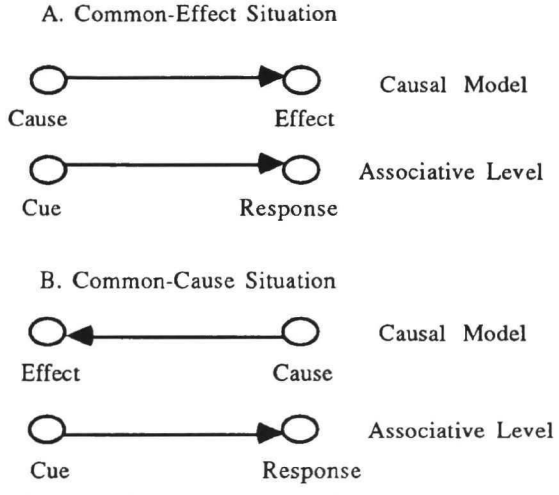

Figure 1. (A) Common-effect situation, in which causal cues are used to predict a potential effect; (B) Common-cause situation, in which presented effects serve as cues to diagnose a potential cause.

The lower halves of both Figure 1A and B show an associationistic representation of the learning situations: Cues are presented first, and the task is to learn to associate them with the correct responses. The corresponding upper halves show how the tasks map to a causal account. In a "common-effect" situation (Figure 1A), the cues represent causes and the responses represent decisions about a predicted effect. Since according to our world knowledge causes always precede effects, the temporal orderings are isomorphic between the causal and associationistic representations of the task. In a "common-cause" situation (Figure 1B), the cues represent effects, and the responses represent decisions about a cause to be diagnosed. The temporal ordering is reversed relative to a common-effect situation, and the mapping to the associationistic description is also reversed. By comparing learning in these two types of situations, this design allows us to disentangle predictions based on associative accounts from those derived from assumptions about causal models for induction. In both tasks, cues have to be associated with the required responses. Thus if subjects treat both tasks as associative multiple-cue learning, then both should yield identical patterns of learning. If, however, subjects represent the two situations in terms of causal models, the two tasks will differ in psychologically important ways, and the learning patterns should reflect these differences.

Experiment 1

Method

A multiple-cue learning task was used in this experiment. Subjects were handed index cards one at a time, with each giving a description of a fictitious person. Subjects were asked to give a "yes-no" response, classifying the cards either as positive or as negative cases. Immediately after every response subjects were told if their judgment was correct or incorrrect. The subjects were trained until they reached a learning criterion (two cycles through the eight basic cases without error) or until they received an upper limit of learning trials.

The descriptions on the index cards consisted of three binary values of dimensional features: weight, pallor, and perspiration. The fictitious persons had either high (e.g., anorexic) or low (e.g., underweight) intensity values on each of these dimensions. The eight possible cases were arranged either in a linearly separable or in a non-linearly separable, correlated fashion (see Figure 2). Similar structures have previously been investigated by Wattenmaker et al. (1986) and by Shepard, Hovland, and Jenkins (1961).

	+				-		
	Case	Dimensions			Case	Dimensions		
		1	2	3		1	2	3
Linearly Separable	1.	H	H	H	5.	H	L	L
	2.	H	H	L	6.	L	L	H
	3.	H	L	H	7.	L	H	L
	4.	L	H	H	8.	L	L	L
Correlated	1.	H	H	H	5.	L	H	H
	2.	H	L	H	6.	H	L	L
	3.	L	H	L	7.	L	L	H
	4.	L	L	L	8.	H	H	L

Figure 2: Structure of item sets used in Experiment 1

The positive set corresponds to a correct "yes" response, and the negative set to a correct "no" response. In the linearly separable arrangement high values of the dimensions are more typical for the positive set, and low values for the negative set. For both sets, each dimension has one exceptional value so that the dimensional values are only probabilistically related to the sets. However, a simple

linear rule distinguishes the two sets. If a person has at least two out of three high values on the three dimensions, then this person belongs to the positive set. This structure does not require hidden layers or configural nodes in a connectionist learning network.

In the correlated, non-linearly separable condition, neither high nor low values are more or less typical for the positive or negative set. For each dimension, there are two persons with high values and two with low values in each set. There is therefore no linear rule to separate the two sets. The only way to distinguish the two sets is to notice the positive correlation between the first and the third dimension in the positive set, and the negative correlation in the negative set. The middle dimension is irrelevant for the classification. This task, which is formally equivalent to learning an "exclusive-or" structure, requires configural nodes or hidden layers in connectionist networks.

This linear-separability factor was crossed with a second factor involving manipulation of the causal structure imposed on the learning task. In the "common-cause" condition, subjects were told that they are going to learn about a disease that is caused by a virus, which could be more or less intense. In this condition the virus plays the role of a common cause that simultaneously affects the symptoms. The cues that subjects saw on the index cards thus correspond to effects of a common cause. This causal model naturally predicts a "spurious correlation" between the effects: A high-intensity virus should yield high-intensity effects, whereas a low-intensity virus should yield low-intensity effects. This situation in fact corresponds to the correlated condition; accordingly, we predicted that this condition should be particularly easy for subjects who received a cover story consistent with the common-cause model.

In a second causal context, the "common-effect" condition, the causal directions were reversed. Now the subjects were told that an experiment on social cognition had been conducted. In this experiment it was found that the appearance of some people produces a new emotional response in their observers. Here the cues on the index cards correspond to potential causes of a common effect. The subjects' task was to learn to predict which person elicits an emotional response in an observer. This emotional response might vary in intensity. Common-effect structures do not imply correlations among the causes. Learning correlated causes amounts to learning a disordinal interaction, whereas the linear condition corresponds to a causal model with three main effects. Given the preference people have for linear as opposed to configural causal structures, the linearly separable task should be relatively easy to learn (see Dawes, 1982).

It is important to note that although subjects were informed that the cause (common-cause condition) or effect (common-effect condition) could vary in intensity, no feedback about the intensity level of the outcome factor was ever provided. Rather, subjects were only told whether the outcome was obtained, regardless of its intensity.

To summarize, if subjects learn according to the accounts of associative learning theories (e.g., connectionist models with hidden layers or configural nodes), the different causal structures imposed on the task should not matter. Subjects across the two causal conditions see identical cues, and are required to learn identical cue-response mappings. However, if subjects are sensitive to the different structural implications of the two causal models, their learning rates for the linear and correlated condition should vary across the two causal cover stories.

Results

Figure 3 shows the results based on 40 UCLA undergraduates who served as subjects. The mean number of errors made prior to the subject reaching the learning criterion was used as an indicator of learning difficulty. As predicted, the causal cover story interacted with the structure of the item set, $F(1, 36) = 7.48$, $p < .025$. The correlated condition was easier to learn in the disease context, in which a correlation naturally falls out of a common-cause structure. In contrast, in the emotional-response condition the linearly-separable item set was easier to learn than the correlated set, as would be expected if people find main-effect models simpler to learn than causal interactions. Overall, the linear condition was learned with fewer errors than was the correlated condition, $F(1,36) = 5.78$, $p < .025$. The results of Experiment 1 thus clearly support the claim that subjects were using causal models during learning, rather than simply trying to associate the presented cues with the correct responses.

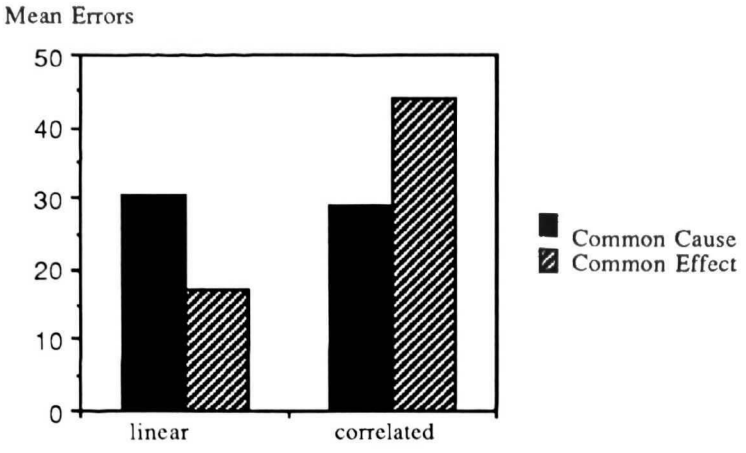

Figure 3. Mean errors prior to reaching criterion as a function of the causal model (common cause vs. common effect) and structure of the item set (linearly separable vs. correlated) in Experiment 1.

Experiment 2

Method

In a second experiment we focused on the correlated condition. In order to better approximate correlations with continuous variables, we used two variables with four intensity levels each in this experiment. We again used weight and pallor as dimensions. The levels of weight were "slightly underweight", "underweight", "seriously underweight", and "anorexic body"; analogous levels were used for pallor. In the positive set these two variables were perfectly positively correlated (values 4 4, 3 3, 2 2, and 1 1, for the four positive items), whereas in the negative set they were negatively correlated (values 4 1, 3 2, 2 3, and 1 4). Note that models like the configural-cue model of Gluck et al. (1989), which introduce separate configural cues for each pairwise feature-value combination, do not capture the monotonicity involved in a correlation of continuous variables. In addition, the number of configural cues required by such models grows exponentially with the number of levels.

In addition to examining learning with more clearly continuous variables, Experiment 2 addressed the question of whether subjects really need explicit information about the fact that the virus (the common cause) may vary in intensity. Even though capturing the positive correlation within the positive set requires the assumption of a continuous common cause, subjects might be able to infer this property of the cause by observing the learning patterns. If the effects are clearly continuous (as was the case for our materials), this may encourage the assumption that the underlying cause is also continuous. Accordingly, half of the subjects received the hint that the common cause might vary in intensity, as in Experiment 1, whereas the other half did not. This hint factor was crossed with the causal context factor, which again consisted of a common-cause and a common-effect condition. Ten subjects served in each of the four conditions.

Results

The results, displayed in Figure 4, replicated the finding that the correlated item set is learned more readily in the common-effect than in the common-cause condition, $F(1,36) = 7.38$, $p < .025$. Omitting the hint that the cause (virus) could vary in intensity did not significantly impair subjects' performance. The impact of causal models on learning correlated item sets thus generalizes to more continuous dimensions.

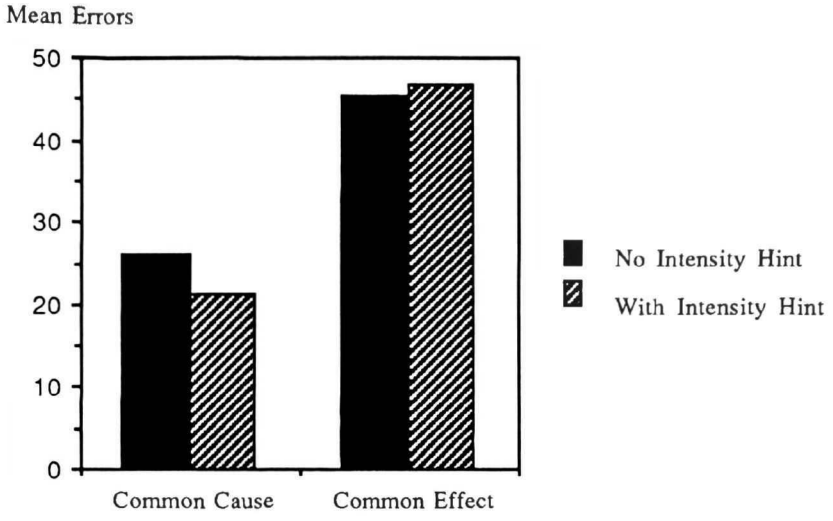

Figure 4. Mean errors prior to reaching criterion as a function of the causal model (common cause vs. common effect) and provision of an intensity hint for a correlated item set in Experiment 2.

Experiment 3

Method

Experiment 3 addresses a restriction that Gluck et al. (1989) imposed on their configural-cue network model. The major advantage of configural-cue networks is that they can learn interactions using simple linear networks with the standard LMS-rule, without requiring backpropagation. Their major problem is that the potential number of configural cues grows exponentially with the number of input cues. Gluck and Bower (1988b) therefore suggest restricting configural cues to pairwise conjunctions. An obvious drawback of this restriction is that such a network is unable to handle problems for which the correct decision requires learning an interaction among three (or more) cues.

	+					-				
Case	Dimensions				Case	Dimensions				
	1	2	3	4		1	2	3	4	
1.	H	H	H	H	5.	H	H	L	H	(L)
2.	H	H	H	L	6.	H	L	H	L	(H)
3.	L	L	L	H	7.	L	H	H	H	(L)
4.	L	L	L	L	8.	L	H	L	L	(H)
					9.	L	L	H	H	(L)
					10.	H	L	L	L	(H)

Figure 5. Structure of item set used in Experiment 3.

In contrast, certain three-way interactions should be learned fairly easily within the context of a common-cause model. Figure 5 shows the structure of the material used in Experiment 3. The positive set was characterized by three correlated dimensions (H H H, or L L L), while the fourth dimension is irrelevant. This type of three-way interaction, like the pairwise interactions used in the correlated conditions in previous experiments, is consistent with a common-cause model in which the cause can vary in intensity. The negative set consisted of the full contrast set with respect to the first three dimensions, so that the subjects indeed had to learn the three-way interaction and could not use

two-way correlations to predict the correct response. In order to keep the negative set small, half of the subjects received the negative cases in which every uneven case has an L-value on the fourth irrelevant dimension, while for the other half these values were reversed. The dimensions and values were the same as those used in Experiment 1, with the addition of two levels of posture as the irrelevant dimension. As in the previous experiment, no hint was given regarding potential intensity variations of the virus. Two groups of subjects differed solely in the causal cover story they received: the disease story (common cause) or the emotional-response story (common effect). Twelve subjects served in each cover-story condition.

Results

The results presented in Figure 6 indicate that the three-way interaction was considerably more difficult to learn than were the correlated item sets within the earlier experiments with pairwise interactions (compare errors for the correlated conditions in Figures 3 and 4). Nonetheless, many subjects were able to attain the criterion of two passes through the items without an error, thus contradicting the implication of Gluck et al.'s (1989) configural-cue model, according to which the task should be unlearnable. In addition, and as in the previous experiments, the common-cause condition yielded a considerably lower error rate than did the common-effect condition, $F(1,22) = 5.40, p < .05$.

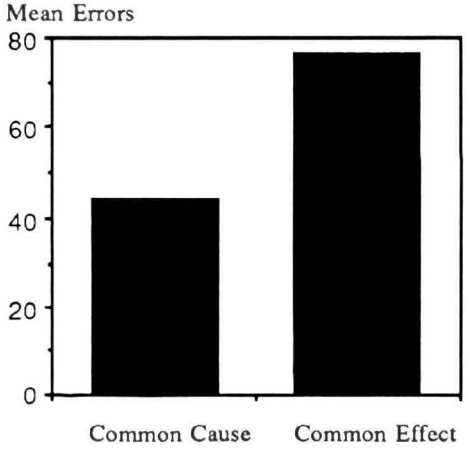

Figure 6. Mean errors prior to reaching criterion as a function of the causal model (common cause vs. common effect) for a correlated item set based on a three-way interaction in Experiment 3.

Discussion

Taken together, the three presented experiments clearly demonstrate the inadequacy of associationistic learning accounts of causal induction as they are embodied in recent connectionist models. Even though the cues and the required responses were identical across the two causal contexts, subjects proved sensitive to the structural implications of the different causal directions implied by the cover stories. Networks that simply code cues on the input layers and responses on the output layer cannot explain such reversal in the relative difficulty of linearly-separable versus correlated item sets, regardless of how they are internally configured.

Our results also demonstrate that causal induction cannot be reduced to associative learning. Associative accounts do not capture the fundamental differences between predictive and diagnostic reasoning (Pearl, 1988). Predictive reasoning requires learning the causal strengths between given causes and potential predicted effects. Once the causal links are learned, information about the presence of a cause allows probabilistic conclusions regarding its likely effects. In diagnostic reasoning, in which causes are inferred from effects, the situation is different. Even with perfect knowledge about cause-effect relationships, effect information is ambiguous with respect to its causes whenever there exists more than one potential cause. Reasoning in this situation requires an inference to the best explanation (Harman, 1986; Thagard, 1989). Different possible theories have to be weighed against each other, and a decision in favor of one or the other theory is based on the fit between the predictions of different theories and the evidence. An analogous approach is taken in research on statistical causal

models as they are embodied in linear structural equations (Bollen, 1989). We are currently modelling the differences between predictive and diagnostic reasoning within a symbolic-connectionist framework, exploring models in which units are interpreted as causes and effects and core links are viewed as causal connections.

Finally, our results are in agreement with many findings demonstrating an overall preference for linear models (e.g., Dawes, 1982; Trabasso & Bower, 1968). Learning linear models puts less strain on information processing because the impact of individual causes is not moderated by the presence of other causes. A number of philosophers have argued that common-cause structures are prevalent in scientific reasoning. Salmon (1984), in particular, argued that theoretical concepts play the role of common causes. Psychologists and philosophers have asked many times what we gain from inferring invisible entities. A possible answer, suggested by the present results, might be that inferred common causes help people to re-represent nonlinear observable structures within a basically linear mental model.

References

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Dawes, R. M. (1982). The robust beauty of improper linear models in decision making. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 391-407). Cambridge: Cambridge University Press.

Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 556-571.

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., Hee, M. R., & Bower, G. H. (1989). A configural-cue network model of animal and human associative learning. In *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Harman, G. (1986). *Change in view.* Cambridge, MA: MIT Press.

Minsky, M., & Papert, S. (1969). *Perceptrons: An introduction to computational geometry.* Cambridge, MA: MIT Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* San Mateo, CA: Morgan Kaufmann.

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.* New York: Appleton-Century-Crofts.

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

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world.* Princeton, NJ: Princeton University Press.

Shanks, D. R. (1990). Connectionism and human learning: Critique of Gluck and Bower (1988). *Journal of Experimental Psychology: General, 119,* 101-104.

Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgment. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory, 21.* New York: Academic Press.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs, 75,* 1-42.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review, 88,* 135-170.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences, 12,* 435-467.

Trabasso, T., & Bower, G. (1968). *Attention in learning: Theory and research.* New York: Wiley.

Wattenmaker, W. D., Dewey, G. I., Murphy, T. D., & Medin, D. L. (1986). Linear separability and concept learning: Context, relational properties, and concept naturalness. *Cognitive Psychology, 18,* 158-194.

Widrow, G., & Hoff, M. E. (1960). Adaptive switching circuits. *Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record, 4,* 96-194.