Alternative Approaches to Causal Induction: The Probabilistic Contrast Versus the Rescorla-Wagner Model

Aaron S. Yarlas
Department of Psychology
U. of California, Los Angeles
Los Angeles, CA 90025
yarlas@psych.ucla.edu

Patricia W. Cheng
Department of Psychology
U. of California, Los Angeles
Los Angeles, CA 90025
cheng@psych.ucla.edu

Keith J. Holyoak
Department of Psychology
U. of California, Los Angeles
Los Angeles, CA 90025
holyoak@psych.ucla.edu

Abstract

Rescorla and Wagner's (1972) model of associative learning (RWM) and Cheng and Novick's (1990, 1991, 1992) Probabilistic Contrast Model (PCM) represent competing approaches to modeling the covariation component of human causal induction. Given certain patterns of environmental inputs to the learner, these models sometimes make contradictory predictions about what will be learned. Some of these situations have been tested in Pavlovian conditioning experiments using animal subjects. We interpret these results according to PCM, and find that they are consistent with the predictions of the model. The current experiment implements similar experimental designs as a causal inference task involving humans as subjects. Two experimental conditions were compared to examine each model's predictions regarding when the extinction of conditioned inhibition will occur. In one condition, the RWM predicts that a previously perceived inhibitory stimulus will be judged as less inhibitory, whereas the PCM predicts that subjects will not change their causal judgments; in the second condition, the two models make the reverse claims. The data provide strong evidence favoring the PCM.

Introduction

Causal induction allows people and other animals to predict and control the environment, a necessary task for survival. How does causal induction occur? What mechanisms are used to induce the causal relation between variables?

A necessary component of causal induction is the evaluation of covariation between a candidate cause and an effect. Two variables that covary tend to be both present or both absent. Two broad classes of models of covariation learning within causal contexts have been proposed. One approach is based on extending associative learning models initially applied to Pavlovian conditioning in animals. The extension is supported by evidence of striking parallels between phenomena observed in studies of causal induction in people and Pavlovian conditioning in animals (e.g., Chapman & Robbins, 1990; Gluck & Bower; 1988; Shanks & Dickinson, 1987; Wasserman, 1990). The most influential associative model has been the Rescorla-Wagner (1972) model of conditioning (RWM), which is a version of the delta rule used to implement learning in many connectionist networks (Sutton & Barto, 1981).

A second approach to modeling causal induction has been influenced by treatments in the philosophical and artificial-intelligence literatures (e.g., Cartwright, 1989; Pearl, 1988). The latter approach has produced models based in part on statistical relations between causes and their effects, as characterized by variants of contingency theory (e.g., Cheng & Novick, 1990; Gallistel, 1990). One formulation of contingency theory, the Probabilistic Contrast Model (PCM) of Cheng and Novick (1990, 1991, 1992), has been extended by Cheng and Holyoak (in press). These alternative theoretical approaches have sparked vigorous debate in the literature (Melz, Cheng, Holyoak & Waldmann, 1993; Shanks, 1991).

The present paper presents a preliminary report of two experimental tests of the RWM and the PCM within causal contexts. For many situations, these two models make similar predictions regarding the causal relations a learner would infer. However, for other situations the two models make diametrically opposite predictions about the causal judgments learners will make. We will report the results of two experimental tests that discriminate between the predictions of the two models.

Rescorla-Wagner Model

The RWM was first proposed to explain various data patterns that had been found in the Pavlovian conditioning literature. The RWM represents the learning of an association by the change in strength of the connection between a conditioned stimulus i (e.g., a flash of light) and an unconditioned stimulus j (e.g., a shock). In addition to the particular stimuli present (e.g., a tone), the stimuli are assumed to include one that represents a context present in every event (e.g., the conditioning cage). In causal terms, each i is a candidate cause, and j is the effect.

Quantitatively, the RWM is represented by the learning rule

$$\Delta V_{ij} = \alpha_i \beta_j \left(\lambda_j - \sum_{k=1}^{n} V_{kj} \right), \qquad (1)$$

where ΔV_{ij} is the change in associative strength between i and j as a result of the current event, α_i and β_j are rate parameters that respectively depend on the salience of i (e.g., the brightness of the light) and j (e.g., the intensity of the shock), and λ_j is the desired output corresponding to the

actual outcome (the presence or absence of the unconditioned stimulus). Typically, if the outcome is present, λ_j is defined as 1; if the outcome is absent, this value is 0. ΣV_{ij}, defined as the sum of the current strengths of associations to j from all n stimuli present in that event, is the actual output of the network predicting the outcome. Learning consists of reducing the discrepancy between the actual outcome (λ_j) and the expected outcome (ΣV_{ij}) until this discrepancy approximates zero.

The strengths that are updated according to Equation 1 are equivalent to weights on the links in a two-layered connectionist network, with the predicting stimuli being represented on the input layer and the predicted outcome on the output layer. An important assumption of the RWM is that if stimulus i is not present during the event, its associative strength remains unchanged (i.e., Equation 1 applies only for those stimuli that are present on a given trial). The equivalent assumption in delta-rule learning is that the strengths of weights from input units with 0 activation are not revised.

Probabilistic Contrast Model

The PCM was proposed by Cheng and Novick (1990, 1991, 1992) to explain the apparent biases that occur when people make causal judgments. This model extends Kelley's (1967) covariation model. Kelley (1967) proposed that people are intuitive scientists, who make causal attributions based on a covariation principle analogous to the analysis of variance. Cheng and Novick (1990) proposed that this principle involves contingency (or contrast), such as that suggested by Jenkins and Ward (1965). Unlike previous contingency models in psychology, however, the PCM assumes that contingency is computed over a *focal set*, which is a set of events a learner considers relevant to the evaluation of the candidate cause. Cheng and Holyoak (in press) reviewed evidence suggesting that when the information is available, the focal set is one in which all alternative plausible causal agents are held constant. That is, the learner often computes the contrast for a candidate cause *conditional* on the constant presence or absence of alternative plausible causes.

The PCM determines the causal relation between a candidate cause and an effect by contrasting the probabilities of the effect being present when the candidate cause is present versus absent within the focal set. A main-effect contrast, Δp_i, which evaluates a candidate cause involving a single factor i, is defined as

$$\Delta p_i = p_i - p_{\bar{i}} \qquad (2)$$

where p_i is the proportion of events for which the effect occurs when factor i is present, and $p_{\bar{i}}$ is the proportion of events for which the effect occurs when factor i is absent (The proportions are estimates of the corresponding conditional probabilities.) If Δp_i is noticeably positive, we perceive i to be an excitatory cause of the effect. If Δp_i is noticeably negative, we perceive i as preventing or inhibiting the effect. If Δp_i is not noticeably different from zero, we perceive i as having no causal relation with the effect.

Conditioned Inhibition

A phenomenon involving multiple stimuli that is predicted by both models is the acquisition of *conditioned inhibition* (Rescorla, 1969). In the standard design, a stimulus A (e.g., a light flash) is first paired with an outcome (e.g., a shock), so that A becomes excitatory. Then, a compound stimulus consisting of A together with a novel stimulus X (e.g., a tone) is repeatedly presented, with the AX combination signaling *absence* of the outcome. Exposure to these events causes X to be perceived as inhibiting the outcome.

The RWM predicts the conditioned inhibition of stimulus X because there is a discrepancy between the actual outcome given the compound AX (shock absent) and the expected outcome based on previous trials with A alone (shock present). This discrepancy leads to a reduction in the strength of X, which must become negative to offset the positive strength of A.

Note that the PCM also predicts the conditioned inhibition of X. The contrast for X -- the difference between the probability of the shock occurring when both the light and tone are present, and the probability of the shock when the light is present and the tone absent (i.e., $P(E|A.X) - P(E|A.\bar{X})$) -- is negative. Thus, both models predict that X will be judged inhibitory, consistent with Rescorla's (1969) finding using animal subjects.

Although both models can account for the acquisition of conditioned inhibition, they make radically different predictions regarding the *extinction* of conditioned inhibition. The extinction of a conditioned inhibiting stimulus (such as X described above) occurs when new information leads to X no longer being perceived as preventative. The RWM predicts that conditioned inhibition will be extinguished by a "direct" procedure, in which a conditioned inhibiting stimulus X is later presented alone with the outcome absent. The RWM predicts that the inhibitory strength of the stimulus will become less negative (eventually reaching asymptote at a strength of zero), as the RWM revises the strength of a stimulus that is present to reduce the discrepancy between the actual and expected outcomes. The model therefore predicts that X will be extinguished as an inhibitor in the direct procedure. In contrast, the PCM predicts that the inhibitory value of X will remain unchanged, as the relevant conditional contrast mentioned earlier, $P(E|A.X) - P(E|A.\bar{X})$, yields an unchanged negative number despite the intervening experience with X in the absence of A. Experiments using this design with animals have yielded support for the predictions of the PCM, in that the direct procedure fails to extinguish conditioned inhibition. Zimmer-Hart and Rescorla (1974) conducted several experiments with rats as subjects, and found that when a previously inhibiting stimulus was presented alone with no outcome, it retained its inhibiting strength in later trials when paired with a novel excitatory stimulus.

The predictions of the two models are reversed for an "indirect" extinction procedure in which a previously excitatory stimulus A, which had been inhibited by a preventative stimulus X, is at a later time no longer paired with the presence of the outcome (i.e., the excitatory power of A is extinguished). Given this information, the RWM

predicts that the inhibitory strength of X will remain unchanged because the RWM cannot update stimuli that are not present, and X is never present during the interval in which the excitor A is extinguished. The PCM, however, predicts that the inhibitory value of X will be attenuated, due to the fact that the relevant conditional contrast, P(E|A.X) – P(E|A.$\overline{X}$), which had been negative when A was excitatory, approaches 0 given the subsequent events (the value of the first term remains at 0, while the value of the second shifts from 1 toward 0). Studies of animal conditioning (Kaplan & Hearst, 1985; Lysle & Fowler, 1985; Miller & Schachtman, 1985) have yielded results consistent with the predictions of the PCM, as conditioned inhibition is extinguished under the indirect procedure.

It thus appears that the PCM provides a more accurate model of Pavlovian conditioning than does the RWM, in that the former model is more congruent with the results of several major animal conditioning studies. To evaluate these alternative models as explanations of causal inference, however, it is necessary to investigate the acquisition and extinction of conditioned inhibition by human subjects who are faced with causal relations. Moreover, previous studies have not directly compared the impact of the direct and indirect procedures on extinction of conditioned inhibition within the same experiment. Accordingly, the present study compares extinction of conditioned inhibition for humans given a causal inference task under both the direct and the indirect procedures.

Method

Subjects

Sixty-one students in undergraduate psychology courses at the University of California, Los Angeles, served as subjects in exchange for course credit.

Design and Procedure

Subjects were given a cover story in which they were told that an outcome (a disease called DSE) was either caused, prevented, or not affected by five candidate causes representing biochemical substances called "endomins," which were said to sometimes be produced by the body. The five candidates were labeled P, Q, R, S and T for subjects (with appropriate counterbalancing); here we will use the more mnemonic labels E_1, E_2, E_3, I, and U, where E indicates an excitatory cause, I an inhibitor, and U a candidate unrelated to the outcome. These candidates were associated with the outcome in specific covariational relationships, which were to be induced by subjects through trial-and-error learning. Candidates E_1, E_2, and E_3 were all causes of the disease, in that when these candidates were present with all other candidates absent, the disease was always present. The disease was always absent when no candidates were present. Candidate I was an inhibitory cause, in that when it was presented in tandem with either cause E_1 or E_2, the disease was no longer present. Candidate U was irrelevant to the disease, in that the disease was always absent (at its baseline) when U was present, just as when U was absent.

In the learning phase of this experiment, all subjects were given a series of learning trials in which they were expected to induce, by making use of feedback, the appropriate causal values for each candidate. All subjects were then tested, using two different measures, for learning of these causal relations. The first measure presented subjects with nine combinations of the various endomins (E_1, E_2, E_3, I, U, E_1 & E_2, E_1 & I, E_2 & I, and E_3 & I) and asked subjects to predict the number of patients out of one hundred who would contract the disease given each of these nine combinations. Note that two of these combinations (I alone, and E_3 & I) were not presented during the learning trials. The second measure presented only the five single endomins, and asked the subject to indicate (by circling one choice) whether each endomin causes, prevents, or has no effect on the disease. The purpose of using two measures was to examine two different types of causal judgments that could be made. The first measure was *implicit*, in that it assessed the inhibitory power of candidate I (the candidate of interest) by subjects' predictions regarding the outcome given that I is presented with a newly paired excitatory cause (E_3). The second measure, in contrast, was *explicit* in that it required subjects to make a direct causal judgment. The RWM and PCM make the same predictions regarding what subjects will learn during the initial learning phase. In particular, I should be judged as inhibitory.

In the extinction phase, which immediately followed, subjects were divided into three groups. In the *control* group, all subjects were given additional trials of some information that had been presented in the initial learning phase; the purpose of this group was to provide a baseline for comparison. In the *direct extinction* condition, which was modeled after the conditions in the Zimmer-Hart and Rescorla (1974) study, subjects were presented with new trials in which the previously inhibitory cause (candidate I) was now presented alone in the absence of the disease. In the *indirect extinction* condition, which was based on the conditions used in the studies of Kaplan and Hearst (1985), Lysle and Fowler (1985), and Miller and Schachtman (1985), subjects received trials in which two previously excitatory causes (E_1 and E_2) were now paired with the absence of the disease. Candidate I was not presented during the indirect extinction phase.

Subjects completed the two measures of causal efficacy in the middle and at the end of the extinction phase. The measurements were taken twice in this phase to determine whether subjects had reached asymptote in their causal judgments after the extinction procedure.

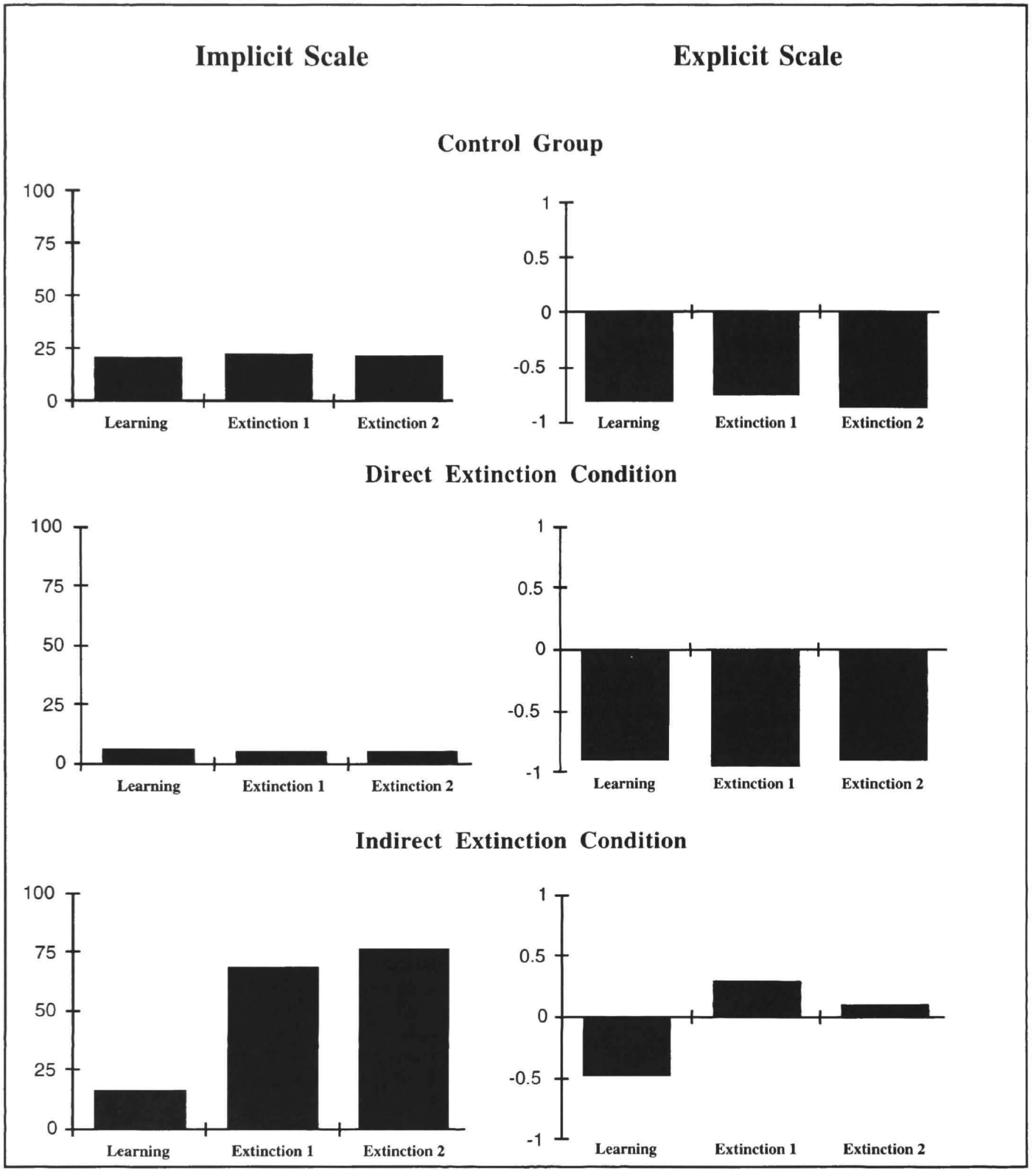

Figure 1: Left column: means for all conditions for the implicit scale (number of patients out of 100 who were judged as having the disease when E_3 & I are present); right column: means for the explicit scale based on causal ratings for candidate I (causes = 1; no effect = 0; prevents = -1).

Results

The bar graphs in the left column of Figure 1 present the results for the implicit measure of the inhibitory power of candidate I across the three test phases (learning, extinction 1, extinction 2): the mean predicted frequency of the disease (maximum of 100) for cases exhibiting the I & E_3 combination. The bar graphs in the right column of Figure 1 depicts the parallel set of results for the explicit measure: the mean causal rating for candidate I on a scale from 1 (causal) to −1 (preventative). The overall pattern of results was qualitatively the same for both causal measures. For each measure, a 3 x 3 analysis of variance with condition (control, direct procedure, indirect procedure) and test phase (learning, first extinction phase, second extinction phase) as independent variables was performed. Both measures yielded a significant interaction between condition and test phase, F (4, 116) = 14.82, $p < .001$ for the implicit measure, and F (4, 116) = 6.83, $p < .001$ for the explicit measure.

Orthogonal contrasts were then used to assess whether the perceived causal power of I changed from the learning phase to the two extinction phases (collapsing across the latter). Neither the control condition nor the direct extinction procedure yielded any change across the test phases on either measure. Both the implicit and the explicit scales revealed that candidate I was perceived as an inhibitor (preventative) of the disease after the initial learning phase, and its inhibitory power remained constant across the later phases. The results for the indirect extinction condition were strikingly different. As the PCM predicts, the inhibitory power of I decreased markedly from learning to extinction, F (1, 58) = 69.83, $p < .001$ for the implicit measure, and F (1, 58) = 25.29, $p < .001$ for the explicit measure.

Discussion

The results from the present study clearly favor the PCM over the RWM as an account of the conditions under which the causal analog of conditioned inhibition can be extinguished. The direct procedure of presenting the inhibitory cause alone in the absence of the effect had no impact at all on its perceived preventative power, whereas the indirect procedure of extinguishing the causal power of a previously excitatory cause essentially eliminated the perceived preventative power of the inhibitory candidate.

These two results each undercut a basic assumption of the RWM and related connectionist learning models. The former undercuts the assumption that the associative strengths of stimuli that are present are revised to reduce the discrepancy between the actual and expected outcomes. Contrary to this assumption, our results show that despite such a discrepancy during a period in which only one stimulus (I) was present, the associative strength of that stimulus was not revised. The latter result undercuts the assumption that only stimuli that are actually present (i.e., have non-zero activation) have their associative strengths revised. Contrary to that assumption, the present experiment shows that the strength of a stimulus (I) was reduced during a period in which it was never presented.

Conversely, these findings support a basic claim of statistical contingency models, namely, that causal judgments are sensitive to the contrast between the probabilities of the effect in the presence versus the absence of a candidate cause, when other causes are held constant.

The present results extend the comparable findings obtained in several classic experiments on Pavlovian conditioning (Kaplan & Hearst, 1985; Lysle & Fowler, 1985; Miller & Schachtman, 1985; Zimmer-Hart & Rescorla, 1974). The present study is the first to compare the direct and indirect extinction procedures within a single experiment. In addition, the present study is the first to test either procedure with human subjects under a causal inference context, rather than animal subjects in Pavlovian conditioning. Our findings reveal that sensitivity to the indirect extinction procedure generalizes from laboratory animals to humans, and from conditioning to explicit, as well as implicit, causal judgments. Our results thus support the contention that the evaluation of covariation in causal contexts is based on sensitivity to statistical information that goes beyond the kind of information implicitly tallied by associationistic models of learning.

Acknowledgments

This material is based upon work supported under a National Science Foundation Graduate Research Fellowship to the first author, and National Science Foundation Grant DBS 9121298 to the second author. We thank Joo-Yong Park and Charles Wharton for their assistance.

References

Cartwright, N. (1989). *Nature's capacities and their measurement.* Oxford: Clarendon Press.

Chapman, G. B., & Robbins, S. I. (1990). Cue interaction in human contingency judgment. *Memory & Cognition, 18,* 537-545.

Cheng, P. W., & Holyoak, K. J. (in press). Complex adaptive systems as intuitive statisticians: Causality, contingency, and prediction. In J.-A. Meyer & H. Roitblat (Eds.), *Comparative approaches to cognition.* Cambridge, MA: MIT Press.

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology, 58,* 545-567.

Cheng, P. W., & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition, 40,* 83-120.

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review, 99,* 365-382.

Gallistel, C. R. (1990). *The organization of learning.* Cambridge, MA: MIT Press.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117,* 227-247.

Jenkins, H., & Ward, W. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs, 79,* 1-17.

Kaplan, P. S., & Hearst, E. (1985). Contextual control and excitatory versus inhibitory learning: Studies of extinction, reinstatement, and interference. In P. D.

Balsam & A. Tomie (Eds.), *Context and learning* (pp. 195-224). Hillsdale, NJ: Erlbaum.

Kelley, H.H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska symposium on motivation* (Vol. 15, pp. 192-238). Lincoln: University of Nebraska Press.

Lysle, D. T., & Fowler, H. (1985). Inhibition as a "slave" process: Deactivation of conditioned inhibition through extinction of conditioned excitation. *Journal of Experimental Psychology: Animal Behavior Processes, 11*, 71-94.

Melz, E. R., Cheng, P. W., Holyoak, K. J., & Waldmann, M. R. (1993). Cue competition in human categorization: Contingency or the Rescorla-Wagner rule? Comments on Shanks (1991). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 1398-1410.

Miller, R. R., & Schachtman, T. R. (1985). Conditioning context as an associative baseline: Implications for response generation and the nature of conditioned inhibition. In R. R. Miller & N. E. Spear (Eds.), *Information processing in animals: Conditioned inhibition* (pp. 51-88). Hillsdale, NJ: Erlbaum.

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

Rescorla, R. A. (1969). Conditioned inhibition of fear. In W. K. Honing & N. J. Mackintosh (Eds.), *Fundamental issues in associative learning*. Halifax: Dalhousie University Press.

Rescorla, R. A., & Wagner, A. R., (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.

Shanks, D. R. (1991). Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 433-443.

Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgment. In G. H. Bower (Ed.), The psychology of learning and motivation, (Vol. 21, pp. 229-261). New York: Academic Press.

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

Wasserman, E. A. (1990). Attribution of causality to common and distinctive elements of compound stimuli. *Psychological Science, 1*, 298-302.

Zimmer-Hart, C. L., & Rescorla, R. A. (1974). Extinction of Pavlovian conditioned inhibition. *Journal of Comparative and Physiological Psychology, 86*, 837-845.