Journal of Experimental Psychology: General
1992, Vol. 121, No. 2, 222-236

0096-3445/92/$3.00

Predictive and Diagnostic Learning Within Causal Models: Asymmetries in Cue Competition

Michael R. Waldmann
Universität Frankfurt, Frankfurt/Main, Federal Republic of Germany

Keith J. Holyoak
University of California, Los Angeles

Several researchers have recently claimed that higher order types of learning, such as categorization and causal induction, can be reduced to lower order associative learning. These claims are based in part on reports of cue competition in higher order learning, apparently analogous to *blocking* in classical conditioning. Three experiments are reported in which subjects had to learn to respond on the basis of cues that were defined either as possible causes of a common effect (predictive learning) or as possible effects of a common cause (diagnostic learning). The results indicate that diagnostic and predictive reasoning, far from being identical as predicted by associationistic models, are not even symmetrical. Although cue competition occurs among multiple possible causes during predictive learning, multiple possible effects need not compete during diagnostic learning. The results favor a causal-model theory.

Tasks as different as classical conditioning, category learning, and causal induction can be viewed as examples of multiple-cue contingency learning. In each of these tasks, a number of cues, which might represent conditional stimuli, features, or causes, are combined to elicit a response. Because of this apparent formal similarity between different types of multiple-cue learning situations, it is tempting to postulate a common learning mechanism. Indeed, a number of researchers have recently claimed that higher order types of learning, such as categorization and causal induction, can be explained by principles that govern lower order learning in animals, such as classical conditioning. Gluck and Bower (1988), for example, suggested that adaptive associative networks can provide powerful models of human categorization. These connectionist networks 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 of the network are computed by a linear function of the weighted cues. The weights are learned 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 showed that a simple model of this sort compares favorably with other models of human categorization (see also Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; for a critique see Shanks, 1990a, 1990b). Because the LMS rule is formally equivalent to Rescorla and Wagner's (1972) theory of classical conditioning (Sutton & Barto, 1981), these findings suggest that categorization can be viewed as a special case of associative learning. Similarly, Shanks and Dickinson (1987) argued that learning of causal relationships can be reduced to associative learning (see also Wasserman, 1990). In the associative framework, cues typically correspond to potential causes, and responses correspond to predictions of potential effects. Weights representing the strengths of the relationships between causes and effects are learned in an incremental fashion.

Cue Competition in Associative Models of Multiple-Cue Contingency Learning

All modern associative learning theories emphasize the competitiveness of cues; indeed, cue competition can be viewed as the single most important feature of current associative learning theories (see Gallistel, 1990). The classic evidence for cue competition involves the phenomenon of blocking, first observed by Kamin (1969) in experiments on aversive conditioning in rats. Such blocking experiments typically consist of two learning phases. In Phase 1, a rat learns to associate an initial conditioned stimulus (CS_1), for example, a tone, with an unconditioned stimulus (US), for example, shock. In Phase 2, the previously conditioned CS_1 (tone) is presented together with a new, redundant CS_2 (e.g., light), and the compound is reinforced by the US. In the critical test phase, the rat sees each CS by itself. As expected, when presented with CS_1 (tone) alone, the rat still shows fear reactions. However, CS_2 (light) typically does not elicit fear reactions, even though the light was constantly paired with the US in Phase 2, and during this period, the shock never occurred in the absence of the light. Learning about the CS_1 seems to have blocked acquisition of associative strength for the CS_2.

Rescorla and Wagner (1972) developed their theory of associative learning to account for blocking and other findings involving cue interactions. Within the Rescorla-Wagner theory, blocking is viewed as the result of the failure of the

This research was supported by a National Institutes of Health Biomedical Research Support Grant to Keith J. Holyoak. A preliminary report of these experiments was presented at the 1990 meeting of the Psychonomic Society in New Orleans, Louisiana. We thank Patricia Cheng, Eric Melz, Frederic Vallee-Tourangeau, and two anonymous reviewers for helpful discussions.

Correspondence concerning this article should be addressed to Michael R. Waldmann, who is now at Psychologisches Institut, Universität Tübingen, Friedrichstraße 21, 7400 Tübingen, Federal Republic of Germany.

second, redundant cue to acquire associative strength. This failure is an inevitable consequence of their proposed learning rule,

$$\Delta V_i = \alpha_i \beta_j (\lambda_j - \Sigma V), \quad (1)$$

which states that the change in associative strength on a trial for each presented cue i, ΔV_i, is proportional to the difference between the outcome concerning US_j that should have been predicted, λ_j, and that predicted by the sum of all current cues, ΣV, weighted by learning rate parameters α_i and β_j that are specific to the particular CS and the US, respectively. When there is no discrepancy between the actual and predicted outcomes, no learning will occur. In blocking experiments, by the end of Phase 1 the animal has already learned that CS_1 predicts the US perfectly (i.e., $\lambda_j - \Sigma V = 0$). Accordingly, no learning will occur for the compound in Phase 2, because changing the (initially 0) associative strength for CS_2 cannot improve the already-perfect predictability of the US. Because of their close formal similarities, the same explanation of blocking is implied by current connectionist learning theories that use the LMS rule or conceptual extensions of it, such as back-propagation algorithms (Rumelhart, Hinton, & Williams, 1986).

Although Rescorla and Wagner's (1972) theory remains influential, other associative accounts of blocking phenomena have also been proposed. Mackintosh (1975) suggested that blocking results from decreases in the associability of cues. In his theory, the learning rate parameter associated with CS_2 declines to the extent that the cue is a worse predictor of the outcome of each compound trial than CS_1. Because CS_2 starts with low associative strength at the beginning of Phase 2, its predictive value compares unfavorably with CS_1, which leads to a decrease in the associability of the new cue during the following trials. In contrast, Pearce and Hall (1980) claimed that a cue loses associability to the extent that its consequences are fully predicted. Because the US is fully predicted by CS_1 as well as by the compound at the beginning of Phase 2, both cues lose their associability, thus preventing further learning. Other theories suggest that blocking is not due to acquisition failure at all, but rather to comparisons between the associative strengths of cues that are made during retrieval of associative knowledge (Miller & Matzel, 1988; Shanks & Dickinson, 1987). Regardless of where the exact locus of blocking is claimed to be, however, all associative theories of multiple-cue contingency learning predict cue competition, with previously acquired strong cues diminishing the impact of later, redundant cues.

Because blocking can be seen as one of the empirical hallmarks of modern associative learning theories, a number of researchers interested in demonstrating the associative underpinnings of causal induction have tried to produce blocking in higher order learning tasks. Shanks and colleagues (Dickinson & Shanks, 1985; Dickinson, Shanks, & Evenden, 1984; Shanks, 1985; Shanks & Dickinson, 1987) conducted experiments in which subjects played a video game requiring them to fire artillery shells at tanks moving through a mine field. One group of subjects went through an observation phase in which they learned that the tanks sometimes explode because they presumably hit a mine in the mine field. Subjects were later allowed to fire at the tanks, after which they rated the effectiveness of their firing. The results indicated that the pretrained group generated lower ratings than did a control group that did not receive preexposure to the mine field. Learning that the mine field was a potent cause of explosions apparently tended to block learning about the artillery shells.

Chapman and Robbins (1990) also demonstrated blocking in predictive causal induction. In their study, subjects had to learn to predict the behavior of a fictitious stock market based on information about individual stocks. Chapman and Robbins used a within-subjects design, in which the trials for each subject were divided into two phases. In the first phase, whenever one stock (P) rose in price, the market rose in value as well. Whenever a second stock (N) rose in price, however, the market failed to increase in value. Thus, P was established as a positive predictor (a leading indicator of change in the stock market), whereas Stock N was nonpredictive. Two other stocks (B and C) never changed during Phase 1. In the second phase, increases in P were paired with increases in B as a second redundant predictor. B never rose by itself, but each time P and B rose together, the market rose. On different trials increases in N were paired with increases in C, and together these cues also predicted the rise of the market value. As predicted, P blocked learning about the contingency between B and market value (as measured by ratings of the predictiveness of the individual cues). Cue C received much higher predictiveness ratings than did B, even though individually B and C were equally perfect predictors. It should be noted, however, that Chapman and Robbins (Experiment 1) obtained only partial blocking of the B cue, whereas without additional assumptions the Rescorla-Wagner model predicted complete blocking in their experimental situation. Wasserman (1990) has also reported evidence of cue-competition effects in human causal induction.

Blocking has rarely been investigated in the context of category learning. Trabasso and Bower (1968) found that subjects disregarded a new redundant cue when they previously had learned to sort stimuli using one single, valid cue. However, their research used artificial concepts (categories of geometric figures) that had no clear causal basis, so the relationship (if any) of their findings to causal induction remains unclear. Gluck and Bower (1988), although they did not use a blocking design, performed a cross-experiment comparison indicating that ratings of the predictiveness of symptoms as cues for a disease were reduced when other highly predictive cues were present. However, Chapman and Robbins (1990) have pointed out that Gluck and Bower's finding may have reflected a change in subjects' use of the rating scale rather than blocking. Shanks (1991) found that subjects' ratings of how strongly associated an individual symptom was with a disease were reduced if a co-occurring symptom was more predictive of the disease for the training examples. The interpretation of these results is unclear, however, because of the vagueness of the question subjects were asked. It might have been interpreted by some subjects as a request to rate the relative predictiveness of symptoms instead of their causal relationship to the diseases.

Other work provides evidence that in some category-learning situations the presence of redundant cues actually produces not blocking but mutual facilitation of learning. Billman (1989), in experiments on category learning and on acquisition of the syntax of an artificial language, found that subjects learned a positive correlation between a pair of cues more readily if each of the cues was involved in other predictive relationships rather than being related only to each other. The conditions that lead or do not lead to cue competition in human category learning are thus not yet well understood.

Learning Within Causal Models

We would like to contrast the associative view of causal learning described earlier with a more mentalistic approach. According to this latter view, people use meaningful world knowledge, often of a highly abstract sort, to guide their learning about new domains. One major example of abstract world knowledge is knowledge about the basic characteristics of causal relations, such as the temporal precedence of causes to their effects. The experiments we report here demonstrate that human causal induction depends on the learner's causal model of the situation and, hence, that causal induction cannot be reduced to associative learning.

The causal-model theory we are advocating embodies three basic assumptions about human causal induction: (a) People have a strong predisposition to learn directed links from causes to their effects, rather than vice versa, even in situations in which they receive effect information prior to cause information; (b) the perceived strength of a causal connection is related to the *contingency* between the possible cause and the effect; and (c) although the links in a causal model are asymmetric (directed from cause to effect), people are nonetheless able to make both *predictive* inferences (from a cause to its likely effects) and *diagnostic* inferences (from effects to their likely causes); furthermore, these two types of inferences have important structural differences. We now consider each of these assumptions in turn.

Directionality of Causal Links

A fundamental psychological constraint on causal reasoning is the assumption that causes must precede their effects. Often, of course, temporal order is directly observable, as when a fire starts and then produces smoke. Even when cues are observed simultaneously or in the reverse of their causal order, however, as when one sees smoke and infers there must be a fire, the natural causal model will still be based on links directed from causes to their effects. The fact that order of observation can be decoupled from temporal precedence within a causal model provides the basis for our experimental dissociations between causal and associative learning. In simple associative models, connections are learned between presented cues (the input layer of an adaptive network) and predicted outcomes (the output layer). In what we term a *predictive causal model* (Figure 1, panel A), the input is interpreted as a cause (e.g., fire), and the output, as an effect (e.g., smoke). In this predictive case—which fits most of the causal-induction experiments in which cue competition has been reported—the temporal order within the causal model coincides with the input–output sequence at the associative level. In a *diagnostic causal model* (Figure 1, panel B), however, temporal order is reversed from the causal model to the associative level. Here the input is interpreted as an effect (e.g., smoke), which is understood to occur after its cause (e.g., fire), even though the effect is presented prior to the cause. In associative models, however, it is the cause that would be assigned to the output layer.

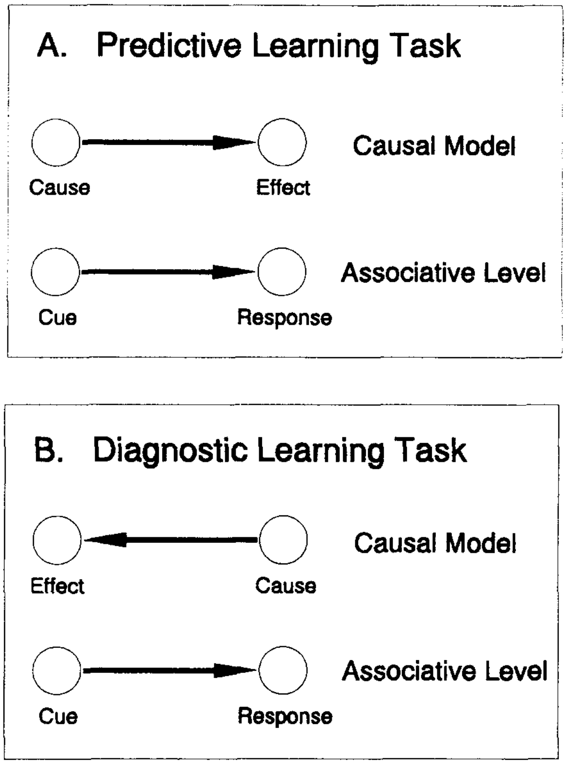


Figure 1. In a predictive learning task (panel A), temporal order within the causal model corresponds to input–output order at the associative level; in a diagnostic learning task (panel B), temporal order within the causal model is opposite to input–output order at the associative level.

The assumption that people preferentially learn links from cause to effect, rather than vice versa, is supported by experimental evidence as well as intuition. For example, Tversky and Kahneman (1980) found that people estimated that it is more likely that a blue-eyed mother will have a blue-eyed daughter than vice versa, even though the corresponding conditional probabilities are necessarily equal (assuming that the prevalence of blue eyes in the population does not differ across generations). Tversky and Kahneman interpreted their findings as evidence that people have directional causal schemata and that people focus on learning relationships between causes (e.g., mother's eye color) and effects (e.g., daughter's eye color). Similarly, Eddy (1982) reviewed evidence that doctors tend to use disease-to-symptom (i.e., cause-to-effect) conditional probabilities even in situations in which symptom-to-disease (i.e., effect-to-cause) conditionals would be normatively appropriate (see also Einhorn & Hogarth, 1986).

Causal Contingency in Multiple-Cue Situations

Causal-model theory provides an alternative to associationist accounts of cue competition, based in part on the statistical concept of *contingency.* It has long been argued that the normative statistical evidence for a causal link between a possible cause and an effect is observing that the proportion of occasions on which the effect is observed is greater in the presence of the possible cause than in its absence. The contingency between a potential cause C and an effect E is given by the so-called Δp rule,

$$\Delta p = p(\mathrm{E}|\mathrm{C}) - p(\mathrm{E}|{\sim}\mathrm{C}), \qquad (2)$$

where ~C signifies the absence of the cause (e.g., Shaklee & Tucker, 1980; Ward & Jenkins, 1965; for a review of research on covariation detection, see Alloy & Tabachnik, 1984). Note that Δp depends not only on the proportion of cases in which the effect and cause co-occur but also on the proportion of cases in which the effect occurs in the absence of the possible cause. Contingency is thus distinct from the simple conditional probability of the effect given the cause. For the case of a single potential cause, the contingency specified by the Δp rule is equivalent to the asymptotic predictive strength of the cue as determined by the Rescorla-Wagner learning rule given in Equation 1, assuming the context is represented as an additional cue that is constantly present (Chapman & Robbins, 1990).

One of the central claims of supporters of associationist theories has been that evidence for cue competition in causal induction contradicts a contingency account (e.g., Chapman & Robbins, 1990; Shanks, 1991; Shanks & Dickinson, 1987). This claim is based on the assumption that the generalization of the Δp rule to multiple-cause situations is simply to derive Δp for each potential cause in isolation (hence failing to predict blocking). However, this assumption is normatively incorrect. In fact, when normatively generalized to the multiple-cause situation, contingency not only provides a qualitative account of the blocking effects reported in the literature on human causal induction but also predicts boundary conditions on when such effects will be observed.

The normative generalization of the contingency concept to situations involving multiple potential causes has a long history in philosophy (Reichenbach, 1956; Salmon, 1984; Suppes, 1970) and artificial intelligence (Pearl, 1988). It has also been proposed as a descriptive model in psychology (Cheng & Novick, 1990, 1992; Kelley, 1967). As indicated by the statistical method of analysis of variance, in multifactorial designs the effects of potential causes (independent variables) have to be cross-tabulated with each other so that interactions among factors can be detected. Although many investigators have reported apparent deviations of human covariation detection from the pattern predicted by the single-factor Δp rule, Cheng and Novick (1990, 1992) presented evidence that when contingency is generalized to the multifactorial case, and the set of events over which subjects are computing contingency is considered, human judgments of causal efficacy are directly related to an unbiased computation of contingency.

Often, in naturalistic situations as well as in the blocking paradigm, one factor has clearly been established as a cause, and the question arises as to whether a second, correlated factor that also has a nonzero contingency (i.e., a nonzero Δp) is also a cause. It is then necessary to test for the *conditional independence* of the effect from the second potential cause (Reichenbach, 1956; Salmon, 1984). That is, does the effect vary with the second factor when the known cause is held constant? If not, then the effect is conditionally independent of the second factor, and any apparent unconditional contingency between the latter factor and the effect can be attributed to the correlation of the second factor with the known cause. The second factor, then, is a spurious rather than a genuine cause. The effect is independent of the second causal factor (C_2) conditional on the first causal factor (C_1) if both

$$p(\mathrm{E}|\mathrm{C}_1.\mathrm{C}_2) - p(\mathrm{E}|\mathrm{C}_1.{\sim}\mathrm{C}_2) = 0 \quad \text{and} \qquad (3a)$$

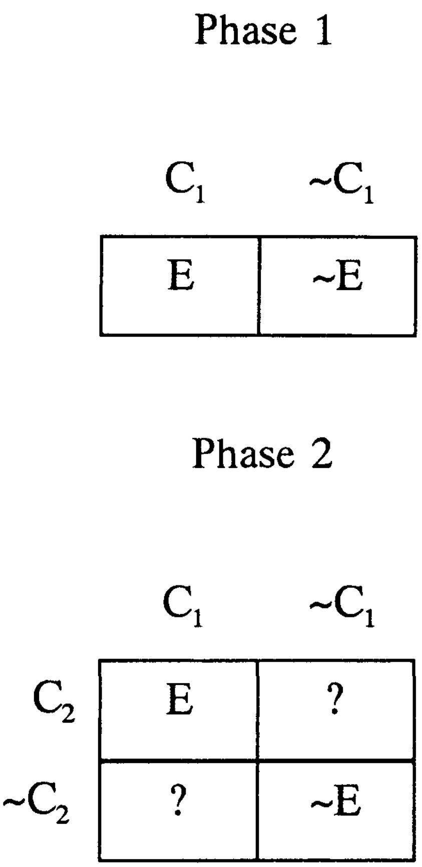
$$p(\mathrm{E}|{\sim}\mathrm{C}_1.\mathrm{C}_2) - p(\mathrm{E}|{\sim}\mathrm{C}_1.{\sim}\mathrm{C}_2) = 0, \qquad (3b)$$

where the dot between C_1 and C_2 denotes *and.*

For example, suppose that smoking (C_1) is known to cause heart disease, and it is found that coffee drinking (C_2), which is correlated with smoking, has an unconditional contingency (i.e., a nonzero Δp) with heart disease (E). If it is then determined that the prevalence of heart disease does not vary among coffee drinkers who smoke (Equation 3a), or among coffee drinkers who do not smoke (Equation 3b), then the conditional-independence criterion will absolve coffee drinking of any causal link to heart disease. This occurs despite the observed unconditional contingency between the two (which is attributable to the correlation between coffee drinking and the genuine cause of heart disease, smoking).

Figure 2 gives a contingency table representation of the blocking design for a predictive causal model of the sort that would be invoked in experiments such as those reported by Shanks and Dickinson (1987) and Chapman and Robbins (1990), in which one causal factor, C_1, is introduced in Phase 1 and then a second, redundant causal factor, C_2, is added in Phase 2. In Phase 1, subjects learn that the effect E occurs when C_1 is present, but not when it is absent. The Δp for C_1 is large (in fact, it equals 1), and hence C_1 is established as a cause. In Phase 2, an incomplete factorial design with missing cell information is created. The new cue, C_2, is constantly paired with the previously established predictive cue, C_1. Although C_2 will also have a large ΔP, it is unclear whether C_2 is an independent cause of E. Accordingly, a test of conditional independence is required.

The blocking design allows a test of Equation 3a. Subjects receive information about the effect of the joint presence of C_1 and C_2 (Phase 2), and the effect of the presence of C_1 in the absence of C_2 (Phase 1). In such a deterministic situation, in which $p(\mathrm{E}|\mathrm{C}_1)$ is 1, the blocking paradigm yields potential overdetermination of the effect. That is, because the effect is already deterministically caused by C_1, another cause cannot possibly increase the probability of the effect. Accordingly, testing Equation 3a cannot establish C_2 as an independent cause, since the ceiling effect renders the test inconclusive. This situation calls for a test of Equation 3b. In particular, it is necessary to check whether the new factor C_2 changes the probability of the effect in the absence of the established cause



Phase 1

C_1	$\sim C_1$
E	~E

Phase 2

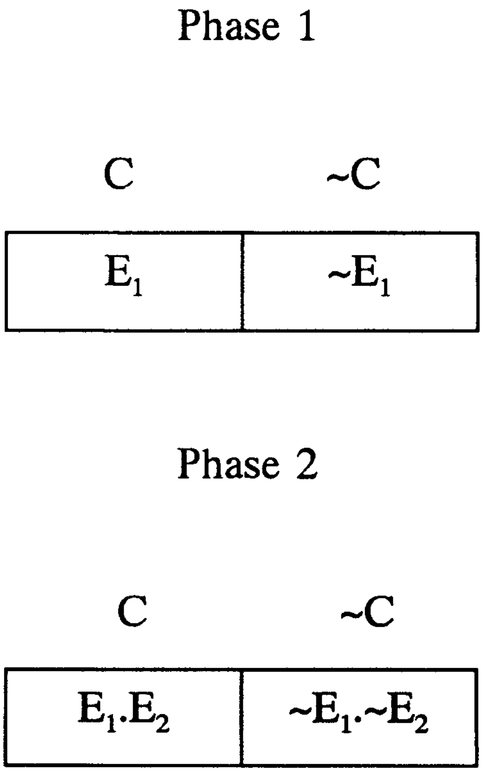
	C_1	$\sim C_1$
C_2	E	?
$\sim C_2$	?	~E

Figure 2. The blocking paradigm within a predictive causal model: Contingency analysis. (C_1 and C_2 are possible causes, and E is the effect. Each cell indicates whether E was observed to be present or absent for some combination of the presence and absence of C_1 and C_2. A question mark indicates a cell for which no information is available because the relevant combination of causes was never presented.)

C_1. However, the cell required for this information ($\sim C_1.C_2$) is missing in blocking designs. It is therefore impossible to determine whether the observed unconditional contingency between the new cue, C_2, and the effect is genuine or spurious. In particular, the crucial missing cell makes it impossible to determine whether the effect is simply overdetermined in Phase 2; whether the second cue, C_2, is correlated with C_1 without having any causal impact by itself (i.e., C_2 is a spurious cause); or whether the two factors produce a causal interaction. This uncertainty should lead to a lowering of confidence in the predictiveness of C_2 (i.e., blocking). Note, however, that unlike the Rescorla-Wagner model, the contingency account predicts that subjects will learn that C_2 is unconditionally correlated with the effect but will be uncertain whether the relationship is causal. Blocking may therefore be incomplete.

A more dramatic decoupling of the predictions of the associative and the causal-model theories is obtained if the blocking paradigm is used with two cues that are interpreted *as multiple possible* effects of a single cause, instead of multiple possible causes of a single effect. This situation, in which the blocking paradigm is used in the context of a diagnostic causal model, is depicted in Figure 3. In Phase 1, the subject now learns that an effect E_1 is obtained when cause C is present, but not when C is absent. Hence, C is established as a cause of E_1. In Phase 2, a new effect, E_2, which is redundant with E_1, is introduced. That is, when C is present the event $E_1.E_2$ occurs, and when C is absent $\sim E_1.\sim E_2$ occurs. Different effects, like different dependent measures obtained in an experiment, do not compete with one another; rather, each effect, as well as any interaction among the effects, provides information about the consequences of the cause. In the blocking paradigm, the learner simply learns a new main-effect contingency between C and E_2: C causes E_2 as well as E_1. Thus, when the causal model is diagnostic rather than predictive, no blocking is expected under the causal-model theory (assuming, as it will prove important in Experiment 2, that only one potential cause of the effects is apparent). In general, causal-model theory predicts a basic difference between the impact of redundancy for causes versus effects: Causes compete, and effects collaborate.

Another way to view the difference in information across predictive and diagnostic contexts, as provided by the blocking paradigm, involves the distinction between absence of observation (specifically, lack of knowledge as to whether a causal connection exists) and observation of absence (knowledge that a causal connection does not exist). If a combination



Phase 1

C	~C
E_1	$\sim E_1$

Phase 2

C	~C
$E_1.E_2$	$\sim E_1.\sim E_2$

Figure 3. The blocking paradigm within a diagnostic causal model: Contingency analysis. (C is a cause, and E_1 and E_2 are possible effects. Each cell indicates which effects are observed to be present or absent in the presence or absence of C. A period denotes *and.*)

of causes is never presented, the learner has no opportunity to observe the consequences of that combination, yielding lack of knowledge concerning its causal efficacy (as in the cells with a question mark in Figure 2). In contrast, if a combination of effects does not occur, the learner is licensed to conclude that none of the presented cause or causes produce the missing effect combination, yielding knowledge that causal efficacy is absent. It follows that whereas the blocking paradigm leads to missing information in a predictive context, and hence uncertainty about the causal status of the redundant cue, there is no parallel uncertainty in the diagnostic context. In the latter case, although observations of the redundant cue, E_2, were not made available in Phase 1, Phase 2 provides all the information required for the learner to decide that C causes E_2 as well as E_1.

In contrast, associative theories provide no basis for distinguishing between predictive and diagnostic causal-induction tasks. Associative theories, regardless of whether they see associative learning as a low-level process (e.g., Gluck & Bower, 1988) or as modification of higher order beliefs (Shanks & Dickinson, 1987), share the fundamental assumption that cues, which are defined as the given information on the basis of which responses are to be predicted, correspond to information obtained prior to outcomes. The semantic distinction between causes and effects cannot be represented in purely associative terms. Associative learning theories, therefore, imply that otherwise identical predictive and diagnostic learning tasks should yield identical learning behavior. In particular, blocking should be obtained regardless of the interpretation of the redundant cues as causes or as effects.

Predictive Versus Diagnostic Inferences

The causal-model theory assumes that although people learn directed contingencies linking causes to their effects, rather than vice versa, they nonetheless can use their knowledge to make both predictive and diagnostic inferences (e.g., Carlson & Dulany, 1988). Crucially, diagnostic inferences, which go from observed effects to possible causes, require mechanisms that do more than assess cause-to-effect contingencies.[1] As Pearl (1988) pointed out, there are important structural differences between diagnostic and predictive reasoning, which neither simple associative models nor contingency computations alone can capture. For example, if you know from past experience that rain causes grass to become wet and to look green, then knowing that it rained licenses you to predict that the grass should be greener and wetter than yesterday. Subsequent observation that the grass is in fact greener than it was yesterday would actually lend additional support to the prediction that the grass should be wet. Thus, multiple effects of a common cause actually support each other in predictive inference (from causes to potential effects). On the other hand, if you see wet grass and have additional evidence that points to a sprinkler as the cause, then rain, as an alternative potential cause of wetness, becomes less plausible than before. Multiple alternative causes thus compete in diagnostic inference (from effects to potential causes). Note that even if the contingency between a cause and its effect is very high, it is not necessarily the case that observing the effect provides strong evidence for the cause (because some alternative cause might actually have produced the effect). Diagnostic reasoning thus requires adjudication among competing causal theories, in addition to knowledge of cause-to-effect contingencies.

Theory competition in diagnostic learning is structurally different from cause competition in predictive learning. If the information available is insufficient to allow testing for interactions or conditional independence, it will be impossible to establish with certainty the status of a redundant potential cause (leading to partial blocking). If the information necessary to calculate independent multifactorial contingencies is provided, however, it is possible for multiple factors to emerge as strong individual causes. Once a factor has been established as an independent cause, predictive inferences can be made on the basis of knowledge of the state of that factor alone. Here the prediction of individual effects does not depend on other effects that also might be predicted by the cause. In contrast, even after the essential contingencies have been learned, diagnostic inferences remain sensitive to knowledge about alternative causes. Even if the contingencies between a cause and each of its multiple effects are equated, the individual effects may be better or worse diagnostic cues for that causal theory depending on the extent to which each of them might also be accounted for by alternative theories. Assessments of the diagnostic implications of effects should therefore prove to be sensitive to background knowledge about potential alternative causal theories. Furthermore, the diagnostic quality of individual effects is not independent of the presence of other effects in diagnostic reasoning. Indeed, the ranking of theories potentially accounting for some initial evidence might actually reverse depending on which additional effects are added to the evidence.

We performed three experiments to determine whether competition among redundant cues varies across predictive and diagnostic learning contexts in ways predicted by our causal-model account but not by associative learning theories.

Experiment 1

The general design of our experiments was to create two learning situations that were identical at the associative level, differing solely in a cover story that established either a predictive (Figure 1, panel A) or a diagnostic (Figure 1, panel B) causal model. Subjects in the predictive and diagnostic conditions received identical cues and had to learn to give identical responses; hence, the two conditions did not differ at the associative level. However, the conditions differed when analyzed in terms of causal models. In the predictive condition, the cues represented causes of a common effect, whereas

[1] Realistic diagnostic reasoning in complex domains such as medicine often involves a combination of what we call *diagnostic* and *predictive* inferences. For example, a doctor might observe both possible effects of a disorder (e.g., abnormal pulse) and possible causes of a disorder (e.g., puncture wounds in the patient's skin) and then make inferences based on both causal directions, which must be integrated to arrive at a diagnosis (e.g., Patel & Groen, 1986).

in the diagnostic condition, they represented effects of a common cause.

In the predictive learning task used in Experiment 1, subjects saw descriptions of features of fictitious persons and had to learn to predict whether people with these features (possible causes) elicit a new kind of emotional response (a common effect) in observers. In contrast, in the diagnostic learning task subjects saw the same features redefined as symptoms (i.e., effects) of a disease caused by a virus (i.e., a common cause). Analyzed within an associationistic framework, both tasks were thus identical. Subjects saw identical cues and had to learn to give identical responses (*yes* or *no*). In the predictive context, the order of presentation of information was isomorphic to the order of events that represents causes and effects: Cues preceded responses just as causes precede their effects. However, this order was reversed in the diagnostic context: Cues here represented events (effects) that occur in the real world after the events (causes) that map onto responses. Experiment 1 thus tested whether subjects treated both tasks as equivalent, as predicted by associative learning theories, or as different with respect to cue competition, as predicted by the causal-model theory. The major hypothesis was that cue competition would only occur when the cues represented causes (predictive condition) but not when they represented effects (diagnostic condition).

A two-phase blocking paradigm was used to examine cue competition. In Phase 1, one of the features (Cue P) used for the descriptions of the persons was established as a perfect, deterministic predictor of the disease or the emotional response. In Phase 2, this feature was paired with a new redundant predictor (Cue R). Even though this new feature individually was also a perfect predictor, it was completely redundant with the previously established predictor. In both the predictive and the diagnostic conditions, subjects were periodically requested to give causal ratings to individual cues. In the predictive context, subjects rated whether each cue caused the emotional response; in the diagnostic context, they were asked to rate whether each cue was affected by the disease. Note that both of these questions (for the diagnostic as well as the predictive condition) were intended to assess the perceived strength of causal relations in the cause–effect (i.e., predictive) direction. Causal-model theory predicted that cue competition would be observed in the predictive context (in which the redundant cue was interpreted as a possible cause) but not in the diagnostic context (in which it was interpreted as a possible effect).

Subjects also saw two additional irrelevant cues, one (Cue C) that was always set to a constant, normal value, and another (Cue U) that varied but was uncorrelated with the target event (as in the design used by Chapman & Robbins, 1990). These two cues allowed additional tests of the adequacy of associative learning theories. The Rescorla-Wagner learning rule, for example, would predict that no associative strength should accrue to the varying, uncorrelated Cue U. Ratings for this cue should therefore be similar to ratings for the blocked, redundant Predictor R in the predictive context. In contrast, the causal-model theory predicts that subjects should be much more certain about the lack of causal status for the U cue than for Cue R. In terms of contingencies, the U cue has a zero main-effect contrast and is excluded by the conditional independence test; hence, subjects can be certain it is not a cause. In contrast, the R cue has a large main-effect contrast, but the requisite information for testing conditional independence is lacking. Thus, subjects simply do not have enough information to determine whether this cue is a cause. We would therefore expect that the average causal rating in the predictive context would be higher for the R cue, for which the causal status is uncertain, than for the U cue, which is clearly not a cause. Such a difference was in fact observed for the predictiveness ratings obtained in Experiment 1 by Chapman and Robbins, who used a task similar to that of the present predictive condition.

We make a further differential prediction for ratings of the constant C cue across the predictive and diagnostic conditions. Associative learning theories predict that ratings for the constant cue should not differ across the two learning conditions (and should in both cases be equal to the ratings given to the redundant and the varying, uncorrelated cues). In contrast, sensitivity to causal direction should manifest itself in distinctly different ratings. In the disease context (diagnostic condition), subjects should learn that the C cue never changes regardless of whether the person has the disease. The contingency between the disease and this cue is thus zero, so it should be clear that the C cue does not represent an effect of the disease. In the emotional-response context (predictive condition), however, the C cue is interpreted as a possible cause. Because this cue never varies, its contingency cannot be calculated (because the cells representing cases in which the C cue is abnormal are missing). Accordingly, subjects should be uncertain whether the C cue is a cause (rather than certain it is not). In contrast, the varying uncorrelated U cue should receive equally low ratings in both the diagnostic and predictive conditions, yielding an interaction between type of nonpredictive cue and condition. Causal-model theory therefore predicts that the C cue should be rated lower in the diagnostic condition (where it is clearly not an effect) than in the predictive condition (where its causal status is in doubt).

Method

Subjects

The subjects were 24 female and male students from the University of Frankfurt/Main, Federal Republic of Germany. They either received course credit or were paid DM 5 for their participation in this experiment. Half of the subjects were randomly assigned to the diagnostic context, and the other half were assigned to the predictive context.

Procedure and Material

Subjects were run in individual sessions. The material was prepared and presented on IBM PC microcomputers, using the software package APT (Poltrock & Foltz, 1988). Before subjects started to work on the computers, they received typed instructions. Subjects in the diagnostic condition were told that they were going to learn about a new disease that is caused by a virus. The virus cannot be observed directly; however, it affects the person's appearance. The subjects were informed that they were going to see descriptions of people on the computer screen and that half of the people had contracted the new disease. If they thought a person had the disease, subjects were to press the *yes* key; otherwise, they were to press the *no* key. They were told that after each decision they would receive corrective feedback. To make sure that subjects paid attention to each feature they were told that they should focus on each because they were expected to rate it periodically as to whether it was affected by the disease. Subjects were also told that speed of responding was unimportant.

Subjects in the predictive condition were told that a series of recent psychological studies had found that some people's appearances elicit a new emotional response in their observers. This emotional response cannot be observed directly, but it can be measured with psychophysical instruments. Except for necessary adjustments to the emotional-response cover story, the rest of the two sets of instructions was identical. In the predictive condition, subjects were informed that they would be asked periodically to rate whether the individual features were causes of the emotional response.

After reading the instructions, the subjects started the learning task. Descriptions of persons were displayed on the computer screen one after another. To emphasize that the descriptions referred to different persons unique initials were displayed for each person being described. In Phase 1 (pretraining), three features were listed in the descriptions, one below the other, in the following sequence (translated here from German): perspiration (Cue C), skin (Cue P), and posture (Cue U). Perspiration was set to *normal* for each person, skin could have the values *pale* or *normal,* and posture varied between *stiff* and *normal.* After the subjects hit one of the two response keys, they received *correct* or *incorrect* as feedback.

Phase 1 consisted of 48 learning trials in which skin was established as the sole predictor. Thus, 24 persons were pale and had the disease (or elicited the emotional response), and the rest had normal skin. Twelve persons in each of these groups had stiff posture, and 12 had normal posture. All persons had normal perspiration. The descriptions were presented in a random order.

After Phase 1, subjects were handed typed sheets with rating instructions and rating scales. In the diagnostic condition, subjects were told that they were to rate each feature individually as to whether it was an effect of the new disease, independent of whether the feature was affected by other diseases they knew about. Subjects were instructed to rate how confident they were in their attributions. The rating scale ranged from *definitely not an effect* (−10) to *definitely an effect* (+10), with 0 meaning *do not know.* Subjects in the predictive condition received similar instructions, except that they were told to rate whether each feature was a cause of the emotional response.

After the ratings, Phase 2 trials began on the computer screen. Subjects were again reminded that they were supposed to form hypotheses and to focus on each individual feature. Phase 2 consisted of two blocks of 48 trials, with ratings after each block. The two blocks had the same trial structure as Phase 1. The only difference was that weight (Cue R) was included as a fourth, redundant feature: Each person who was pale also was underweight, and each person with normal skin also had normal weight. After each block, subjects received the same rating instructions as after Phase 1, except for the inclusion of the fourth feature in the list of features to be rated.

Results and Discussion

Figure 4 depicts the mean ratings for the three features presented in Phase 1 (panel A) and for the four features presented in Phase 2, averaged over the two measurements within Phase 2 (panel B). All subjects but one (who chose +8) rated Cue P +10 after Phase 1. These results demonstrate that Cue P was in fact established as a clear predictor during pretraining, as is required to test for blocking of the redundant cue in Phase 2.

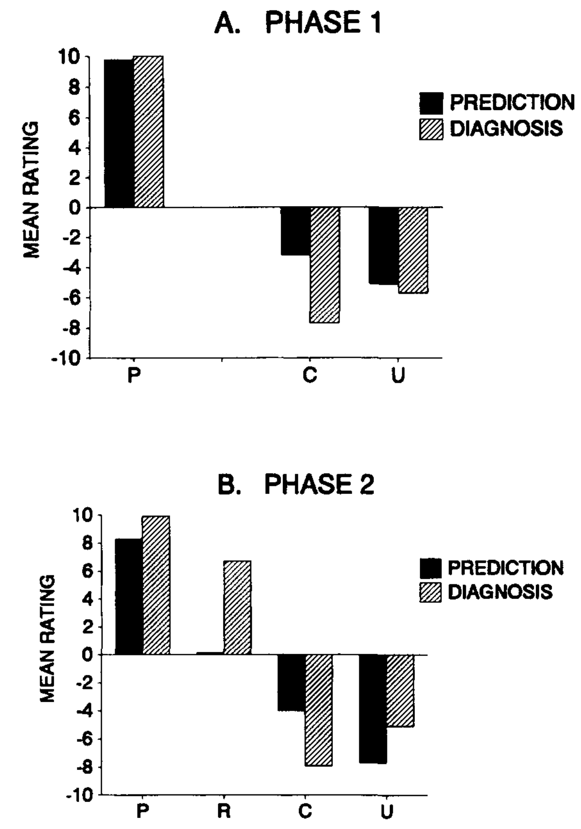


Figure 4. Mean cause ratings (predictive condition) and effect ratings (diagnostic condition) obtained in Phase 1 (panel A) and Phase 2 (panel B) of Experiment 1 for the initial predictive cue (P), the redundant predictive cue (R), the constant uncorrelated cue (C), and the varying uncorrelated cue (U).

The most important analysis involves the comparison of ratings for cues P and R (the redundant cue introduced in Phase 2) across the two causal conditions within Phase 2. An initial analysis including the two blocks of ratings as a factor yielded no significant qualifications of our conclusions in this or subsequent experiments; accordingly, all results were collapsed across the two sets of ratings obtained in Phase 2. A 2 (cues) × 2 (causal conditions) analysis of variance with the two predictive cues (P and R) as a within-subjects factor yielded a reliable effect of causal condition, $F(1, 22) = 9.82$, $MS_e = 13.41$, $p < .01$, and of cue, $F(1, 22) = 35.5$, $MS_e = 14.0$, $p < .001$. Most important, the interaction proved highly significant, $F(1, 22) = 8.94$, $MS_e = 14.0$, $p < .01$. As can be seen in Figure 2, panel B, only in the predictive condition did the ratings for the redundant R cue indicate blocking by the previously acquired P cue, $F(1, 11) = 67.5$, $MS_e = 16.6$, $p < .001$. In the diagnostic condition, the difference between the P cue and the R cue fell short of statistical significance, $F(1, 11) = 3.14$, $MS_e = 39.4$, $p > .10$. The two predictive cues competed only when they represented causes, not when they represented effects.

In a further analysis, the two irrelevant cues were compared with the R cue within the predictive condition. Ratings were again averaged over the two measurements obtained in Phase 2. Planned comparisons revealed that the R cue received a significantly higher causal rating than did the two uncorrelated cues, $F(1, 11) = 21.6$, $MS_e = 13.5$, $p < .01$. This finding indicates that subjects differentiated between the R cue, for which they simply lacked conclusive positive evidence for attributing causal status, and uncorrelated cues for which there was no evidence supporting a causal status. This interpretation is further supported by anecdotal reports provided by the subjects in the predictive condition, most of whom mentioned spontaneously that they could not really make a firm decision about the weight cue (R) because they never saw an underweight person with normal skin. In contrast, without additional assumptions the Rescorla–Wagner model would predict equal ratings for the R cue and for the two uncorrelated cues.

Finally, Phase 2 ratings for the two irrelevant cues were analyzed in a 2 (cues) × 2 (causal conditions) analysis of variance, with the cues constituting a within-subjects factor. This analysis yielded a reliable interaction effect, $F(1, 22) = 5.45$, $MS_e = 23.2$, $p < .05$. Further analyses indicated that the interaction was mainly due to the significantly lower ratings given to the constant C cue in the diagnostic than the predictive condition, $F(1, 22) = 5.74$, $MS_e = 15.7$, $p < .05$. Ratings for the varying, uncorrelated U cue did not yield a significant difference across the two conditions. The same pattern is apparent in the results obtained in Phase 1, where only the ratings for the C cue varied across the two conditions, $F(1, 22) = 5.12$, $MS_e = 20.8$, $p < .05$. This pattern is predicted by the causal-model theory. The U cue should be seen as causally irrelevant in both the diagnostic and the predictive contexts, because the corresponding main-effect contrasts were zero in both cases. In contrast, subjects should have been more confident that the C cue was irrelevant in the diagnostic condition, where it was a noncontingent effect, than in the predictive condition, where it was a potential cause for which neither the main-effect contrast nor the conditional-independence test could be computed, due to missing cells in the contingency table.

To summarize, the results of Experiment 1 clearly demonstrate that causal induction is guided by causal models that differentiate between predictive and diagnostic learning tasks, even when the cue response relationships are identical at the associative level.

Experiment 2

In Experiment 1, we demonstrated that the blocking effect interacted with causal directionality when cause ratings obtained in the predictive condition were compared with effect ratings obtained in the diagnostic condition, providing evidence that subjects in the diagnostic condition were able to form cause-to-effect representations and to assess the strength of effects conditionalized on a hypothetical cause. In true diagnostic reasoning, however, people must reason in the opposite direction, from given effects to hypothetical causes. Categorization studies based on fictitious diseases typically have asked subjects to use symptoms to predict diseases (e.g., Gluck & Bower, 1988), a task that would seem to involve diagnostic inference. In Experiments 2 and 3, we asked subjects to rate the predictiveness of cues (Chapman & Robbins, 1990), regardless of whether the cues represented causes or effects. Thus, both predictive and diagnostic ratings were to be based on strength assessed in the cue-to-response direction as defined at the associative level.

In Experiment 2, we used a learning task identical to that of Experiment 1, with the sole difference that subjects were instructed to give predictiveness ratings in both conditions. Thus, subjects in the diagnostic condition of Experiment 2 were asked questions intended to elicit diagnostic inferences (from effects to a hypothetical cause), whereas the comparable subjects in Experiment 1 were asked questions intended to elicit predictive inferences (from a given cause to its effects).

Method

Subjects

Subjects were 20 female and male students from the University of Frankfurt/Main who either received course credit or were paid DM 5 for their participation. Half of the subjects were randomly assigned to the diagnostic condition, and the other half were assigned to the predictive condition.

Procedures and Material

The procedures and material in Experiment 2 were virtually identical to those used in Experiment 1. The only difference was that all subjects received instructions to rate the predictiveness of each cue individually. As a clarification, subjects were told that they were expected to rate each cue independent of the other cues they had seen together with it. Because no negative contingencies were presented, the rating scale ranged from *not a predictor* (0) to *perfect predictor* (+10).

Results and Discussion

Figure 5 presents the subject's mean ratings, calculated in the same manner as those for Experiment 1 (see Figure 4). In Phase 1 (Figure 5, panel A), subjects learned that cue P was the single valid predictor.

The most interesting analysis involves a comparison of the two predictive cues in Phase 2 (Figure 5, panel B). As in Experiment 1, the two measurements within Phase 2 were averaged for this analysis. A 2 (cues) × 2 (causal contexts) analysis of variance, with the P and R cues constituting the two levels of a within-subjects factor, revealed that the cue that was valid in both Phase 1 and Phase 2 (P) was rated as significantly more predictive than the redundant valid cue introduced in Phase 2 (R), $F(1, 18) = 15.4$, $MS_e = 10.6$, $p < .01$. In contrast to Experiment 1, the interaction was clearly nonsignificant ($F < 1$).

At first blush, the clear attenuation of predictiveness ratings in the diagnostic context poses a puzzle. This aspect of the results is exactly what would be predicted by an associative theory of causal induction. As we pointed out in the intro-

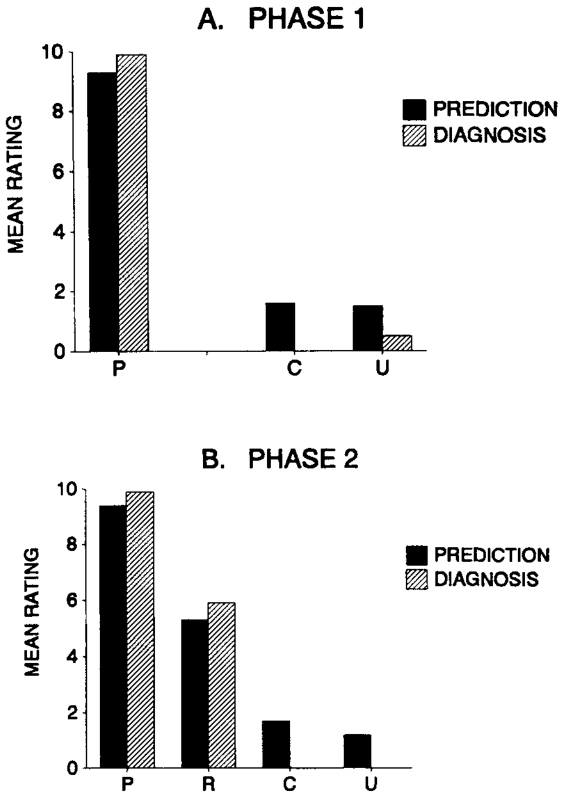


Figure 5. Mean predictiveness ratings for predictive and diagnostic conditions obtained in Phase 1 (panel A) and Phase 2 (panel B) of Experiment 2 for the initial predictive cue (P), the redundant predictive cue (R), the constant uncorrelated cue (C), and the varying uncorrelated cue (U).

duction, however, low predictiveness ratings for single cues could be due to theory competition as opposed to cue competition. The attenuation of the ratings in the diagnostic condition of Experiment 2 can actually be interpreted as a normative result of diagnostic reasoning. Let us assume that subjects in the diagnostic conditions of both Experiments 1 and 2 learned a causal structure in which the virus was a common cause of both being pale and being underweight. In Experiment 1, subjects were asked to rate whether each of these two cues represented effects of the common cause. As would be expected from a normative account, effects did not compete when subjects reasoned from a hypothetical cause to its effects, and so subjects gave high effect ratings to both cues.

The subjects' situation was quite different in Experiment 2, however: They had to reason in the reverse direction, from given effects to hypothetical causes. No attenuation would be expected if there were only one possible theory explaining the evidence; however, this is seldom the case in realistic diagnostic tasks. There is potential competition if multiple causal theories might account for a given effect. More specifically, if subjects bring to bear prior knowledge of alternative possible causes of an abnormal body sign, they may lower their rated predictiveness of the sign as an indicant of the fictitious disease being taught in the experiment. Thus, even though subjects learned in Phase 2 that being underweight (Cue R) is an effect of the virus, they presumably also knew that there are many other reasons a person might be underweight. These alternative causes may have competed with the newly acquired cause as a possible explanation of the single cue *underweight,* causing subjects to give the single cue a relatively low predictiveness rating. The fact that the R cue was introduced only in Phase 2, and was always presented in compound with Cue P, may have made it more likely to evoke extraexperimental knowledge of alternative causes than was the case for Cue P, which had been established as a separable effect of the virus from the outset of the experiment. If this interpretation of the apparent blocking observed in the diagnostic condition of Experiment 2 is correct, then such attenuation of predictiveness ratings should be eliminated if prior knowledge does not provide alternative, extraexperimental causes. We tested this prediction in Experiment 3.

As in Experiment 1, and contrary to the apparent prediction of the Rescorla-Wagner model, in both causal conditions the redundant Cue R received much higher ratings than did either of the uncorrelated cues. As in Experiment 1, the constant C cue tended to yield higher ratings in the predictive condition than in the diagnostic condition. However, because both irrelevant cues were rated 0 by all subjects in Phase 2 of the diagnostic condition (thus yielding 0 variances for these cells), meaningful statistical analyses could not be performed on ratings for these cues.

Experiment 3

The results for the diagnostic condition in Experiment 2 suggested that subjects may tend to consider prior known causes of familiar abnormal symptoms when evaluating the predictiveness of individual symptoms. To eliminate the possible effects of prior knowledge, in Experiment 3 we investigated predictive and diagnostic causal induction in a relatively unfamiliar context. If we are correct in our suggestion that the attenuation of ratings observed in the diagnostic condition of Experiment 2 was a consequence not of cue competition but of true diagnostic reasoning, in which prior knowledge played an important role, then the interaction between ratings and causal context that was obtained in Experiment 1 should surface again in Experiment 3 when prior knowledge is eliminated. In contrast, associative accounts predict cue competition regardless of whether subjects have background knowledge about alternative theories. Note that in experiments on classical conditioning in animals, including those using the blocking paradigm, cues are almost always chosen so as not to have a prior associative history.

Method

Subjects

The subjects were 24 female and male students from the University of Frankfurt/Main, who either received course credit or were paid DM 5 for their participation. Half of the subjects were randomly assigned to the diagnostic condition, and the other half were assigned to the predictive condition.

Procedure and Material

We attempted to construct material that was maximally parallel in structure to that used in Experiments 1 and 2, but for which subjects would lack prior causal knowledge. In Phase 1, subjects received information about the states of three buttons. Button 1 (Cue C) was always constant and set to the value of *off.* Buttons 2 and 3 were *on* in half of the trials and *off* in the rest of the trials. Analogous to the previous experiments, Button 2 (Cue P) was perfectly positively correlated with the correct *yes-no* response, and Button 3 (Cue U) was uncorrelated. As a reminder of the location of the buttons, the label *Room A* was placed next to the three buttons on the computer screen on which they were displayed. In Phase 2, Button 4 (Cue R) was added to the display; it was always on when Button 2 was on and off when Button 2 was off. Button 4 was displayed below the other three buttons and separated by a line. Also, the label *Room B* appeared next to this button, providing additional information to make it clear that Button 4 was located in a different room. As in Experiment 2, subjects were periodically asked to rate the predictiveness of each cue individually, using a rating scale ranging from 0 to 10. The trial structure and the procedure were otherwise identical to those used in the previous experiment.

Predictive condition. In this condition, subjects were told that Peter W. had started to work at a bank in Room A. In the evening, he was expected to switch on the alarm, but unfortunately nobody told him which button turned on the alarm, so he tried out several buttons. The subjects' task in Phase 1 was to learn how to switch on the alarm. Before Phase 2 began, subjects were told that Mary B. had also started to work at the bank on the same day. She was working in a different room (*Room B*), and because she did not know about Peter's attempts, she also tried to switch on the alarm simultaneously. It was mentioned that several buttons were located in her room, but subjects only received information about one button. During Phase 2, Mary, by some accident, only tried Button 4 when Peter switched Button 2, which had been established as the crucial button to activate the alarm.

Diagnostic condition. The cover story in the diagnostic condition was very similar to that used in the predictive condition. The only difference was that in the diagnostic context, Peter and later Mary were trying to figure out whether the alarm was on. They were told that the state of the alarm was signaled by light buttons, which could be either on or off. As no one remembered to tell Peter and Mary which buttons signaled whether the alarm was on, Peter experimentally switched the alarm on and off, and he and Mary checked which signal lights went on or off. As in the predictive condition, subjects saw only the state of the buttons; however, the buttons were redefined as potential effects of a common cause (the alarm). Because the stimuli were identical in both causal conditions, subjects learned in Phase 1 that Button 2 was the crucial signal in Peter's room. In Phase 2, they saw Button 4 (from a different room) as an additional signal for Mary. (In both conditions, two different rooms were introduced to make it more plausible that there may have been more than one crucial button.) The learning task was thus virtually identical in both conditions. In both conditions, subjects received information only about the states of the buttons and had to learn to predict whether the alarm was on or off, making a *yes* or *no* response.

One attractive property of these cover stories is that the cause in the diagnostic condition was identical to the effect in the predictive condition (i.e., the state of the alarm). Thus, the only feature that varied between the two conditions was the causal direction linking the cues and the criterial outcome.

Results and Discussion

Figure 6 presents the mean predictiveness ratings. Subjects clearly learned during Phase 1 that Cue P was a perfect predictor for the alarm, whereas Cues C and U were irrelevant (Figure 6, panel A). The most important analysis, based on the data for Phase 2 (Figure 6, panel B), was a 2 (Cues) × 2 (Causal Conditions) analysis of variance, with ratings for Cues P and R (averaged over the two measurements within Phase 2) constituting a within-subjects factor. This analysis yielded significant main effects of the causal condition, $F(1, 22) = 11.3$, $MS_e = 4.90$, $p < .01$, and of the cue factor, $F(1, 22) = 19.4$, $MS_e = 6.14$, $p < .01$. Most crucial, a reliable interaction was obtained, $F(1, 22) = 9.72$, $MS_e = 6.14$, $p < .01$, with the R cue receiving much higher ratings in the diagnostic than in the predictive condition. In fact, significant blocking was obtained only in the predictive condition, $F(1, 11) = 22.5$, $MS_e = 15.4$, $p < .01$. In contrast, no cue competition between the P cue and the R cue was observed in the diagnostic condition, $F(1, 11) = 1.10$, $MS_e = 9.13$, $p > .30$. These findings support our suggestion that the low ratings for the redundant cue obtained in the diagnostic condition of Experiment 2 were due not to blocking in an associationistic sense but rather to competition between a newly learned causal link

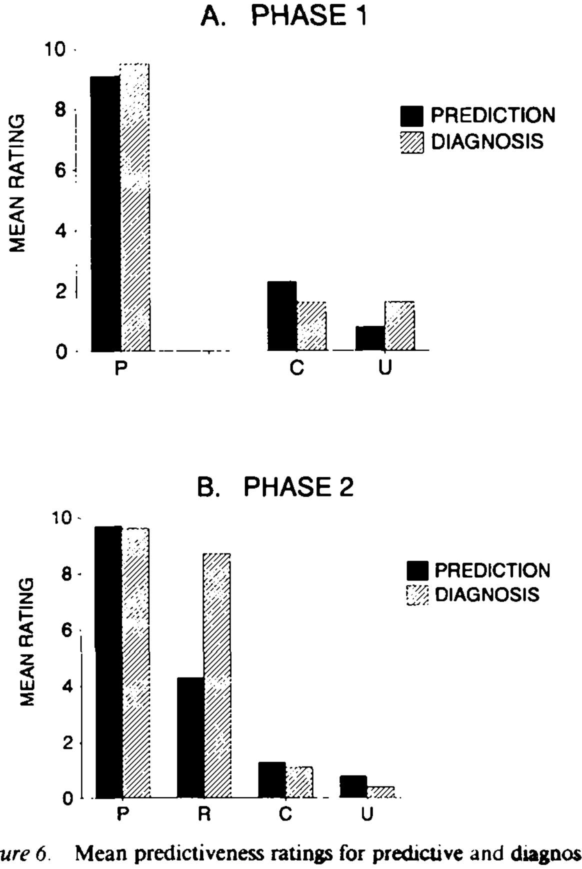


Figure 6. Mean predictiveness ratings for predictive and diagnostic conditions obtained in Phase 1 (panel A) and Phase 2 (panel B) of Experiment 3 for the initial predictive cue (P), the redundant predictive cue (R), the constant uncorrelated cue (C), and the varying uncorrelated cue (U).

and preexperimental causal theories. The unfamiliarity of the material used in the present experiment excluded or reduced the impact of prior causal knowledge; accordingly, the ratings reflect the causal strength connecting the cues and the cause presented in the learning phase. In accord with the prediction of the causal-model approach, cue competion was observed between co-occurring possible causes but not between co-occurring possible effects.

As in the previous experiments, the ratings given to the R cue in the predictive condition, while attenuated, were considerably higher than those given to either of the two uncorrelated cues, $F(1, 11) = 16.6$, $MS_e = 5.17$, $p < .01$. Probably due to cellar effects, analyses involving the two uncorrelated cues did not yield significant results.

General Discussion

Summary

The results of the present study clearly demonstrate that people learn by constructing causal models; furthermore, regardless of the cause–effect order in which the relevant information is presented, these models consist of links directed from causes to their effects. Our results also indicate that people are able to flexibly access this knowledge to generate either predictive or diagnostic inferences. The experiments support the view that diagnostic and predictive reasoning are fundamentally different processes. Far from being identical, as predicted by associationistic accounts of causal induction, the two inferential directions are not even symmetrical. In a predictive learning task, the assignment of causal status to a redundant possible cause is attenuated to some extent, not because of associative blocking per se but rather because of lack of information about cells in the factorial contingency table involving the presence of the second possible cause in the absence of the known cause. In diagnostic learning, by contrast, multiple effects of a common cause do not compete. Knowledge about the presence of a cause licenses the inference of each of its specific effects, independent of other effects that might also emanate from the same cause (Experiment 1). Apparent blocking may occur after diagnostic learning when people use effect cues to predict the cause, not because of competition directly among effects but because of competition among alternative causal theories that might explain the effects, including theories based on prior knowledge (Experiment 2). If alternative causes derived from prior knowledge are ruled out, so that only one possible cause of the observed effects is available, apparent cue competition is eliminated in diagnostic learning (Experiment 3).

Thus, diagnostic inferencing is structurally distinct from predictive inferencing. Diagnosis requires abductive reasoning, or inference to the best explanation, a process that has to deal with potential theory competition (Harman, 1986; Pearl, 1988; Peng & Reggia, 1990; Thagard, 1989). Even deterministic effects of a particular cause may not be perfectly valid diagnostic cues for it, depending on how the cues are related to competing theories. Furthermore, the diagnostic value of a single effect need not be independent of other co-occurring effects. Single cues are rarely sufficient for generating unambiguous diagnoses. Indeed, the relative plausibility of two theories potentially accounting for the evidence might reverse, depending on which second effect cue is added to an initial cue.

Given the dependence of diagnosis on multiple effect cues, we might expect that people will be more sensitive to patterns of cues in diagnostic as opposed to predictive learning. We (Waldmann & Holyoak, 1990) recently presented evidence that supports this prediction. In these experiments, we showed that subjects are more or less sensitive to within-category correlations, depending on the causal context they are provided. Correlated evidence is a natural consequence of common-cause contexts, which are typical for diagnostic task domains and for which use of patterns of cues provides an effective means to constrain the search and evaluation space. The situation is very different in predictive reasoning, in which sensitivity to patterns of causes would require learning about interacting causes. A number of empirical studies have demonstrated that people have a preference for linear arrangements of cues (Dawes, 1982; Mellers, 1980). Given the complexities of the computation of interaction contrasts in a multifactorial contingency table, the cognitive basis for this preference seems obvious. Unlike the case of diagnostic reasoning, where the use of patterns is crucial for constraining diagnoses even for competing linear theories, predictive reasoning favors inferring common effects from their causes in a monotonic fashion. This is only possible if the causes combine in a linear rather than an interactive fashion to produce their effects.

Causal Models and Connectionist Learning Theories

The results presented in this article clearly refute connectionist learning theories that subscribe to an associationistic representation of events as cues and responses. One response to the evidence presented here and elsewhere that learning makes use of directed causal models (e.g., Waldmann & Holyoak, 1990) is to develop network models that embody the assumptions of causal-model theory. For example, cues that are interpreted as causes could be represented on the input layer, and cues that are interpreted as effects could be represented on the output layer, regardless of the temporal order in which the cues are received. Links would then represent directed causal relations (see Pearl, 1988, and Peng & Reggia, 1990, for causal network theories with this general character). Of course, the generation of responses would become much more complicated than in standard associative networks because the responses could not simply be elicited by the cues presented first. Rather, in diagnostic tasks the inputs would have to be assigned to the output level of the causal network. The observable pattern on the output layer would then have to be interpreted as being caused by an unseen input, which would have to be induced by diagnostic learning. Although it may be possible to develop a connectionist model of causal induction along these lines, it should be clear that a number of problems must be solved to account for diagnostic learning. Diagnostic and predictive learning would have to be modeled by structurally different network models. Such models would go far beyond simple associationism and, in fact, would instantiate the causal-model approach we advocate.

Causal Models and Categorization

According to the causal-model approach, categorization is closely linked to causal induction. A number of philosophers and psychologists have recently argued against similarity-based categorization theories, which fail to posit a role for causal knowledge in categorization. Murphy and Medin (1985), for example, claimed that the coherence of real-world concepts is derived from theoretical knowledge they embody (also see Medin, 1989). Putnam (1975) argued that natural-kind concepts are organized around identities of theoretically induced hidden commonalities, which he distinguished from observable features that only provide indirect cues to conceptual identity (also see Gelman, 1988). Keil (1989) extended this view, arguing that conceptual knowledge frequently embodies information about underlying causal as opposed to associative relationships. The relevance of causal knowledge is particularly plausible in tasks involving disease classification, which are frequently used in categorization research (e.g., Gluck & Bower, 1988). Most diseases are defined by hidden causes, such as viruses, which are not directly observable. Here, categorization has to rely on indirect indicators, such as symptoms, which are probabilistic effects of the hidden causes. In such cases, category learning can be viewed as a diagnostic learning task, in which potential causes have to be inferred from given effects.

However, not all categories seem to conform to the causal structure of diagnostic categories. Artifact categories and ad hoc categories frequently seem to be based on features that correspond to causes of a common effect (see Barsalou, 1983; Gelman, 1988; Keil, 1989). For example, *refrigerator* is defined by features that are causally linked to achievement of its intended function of keeping food or other items cool. Such categories are analogous to a class of persons that cause a specific emotional response, the cause category acquired in the predictive condition of Experiments 1 and 2 in this article. Our experiments thus imply that learning about such predictive categories differs from the induction processes involved in learning about diagnostic categories. A causal-model analysis may help to explain the apparent discrepancies between categorization experiments that yield apparent cue competition (e.g., Trabasso & Bower, 1968) and those that actually yield mutual facilitation among correlated cues (Billman, 1989). In general, competition among cues should be expected if they are interpreted as causes, the effect of which is identified by the category label. Neutral or facilitatory cue interactions should be expected, however, if cues are interpreted as common effects, the hidden cause of which is identified by the category label. That is, causes compete, but effects collaborate.

Our research is broadly consistent with the notion that the adaptive goal behind categorization is to learn to predict features (Anderson, 1990; Billman, 1989; Holland, Holyoak, Nisbett, & Thagard, 1986). However, the causal-model theory subordinates this goal to the overarching goal of learning about the causal structure of the world. Causal models, which attempt to enforce conditional independence among representations of otherwise correlated events, constrain the vast number of potential predictive relationships about which we might in principle attempt to learn (Pearl, 1988). Causal models therefore provide an effective means to constrain inductive learning. Medin, Wattenmaker, and Hampson (1987) presented empirical evidence that people are more sensitive to predictive relationships between features when prior knowledge provides causal links underlying these regularities.

Possible Changes in Diagnostic Reasoning With Expertise

The experiments in this article involved essentially novice subjects acquiring knowledge about simple causal relationships in predictive or diagnostic contexts. Our results provide support for the assumption that even in the diagnostic task, in which the cues correspond to effects and the responses correspond to causes, people nonetheless represent the links in their mental models in the cause-to-effect direction. Might this pattern of representation change for more expert subjects who learn more complex diagnostic tasks, such as actual medical diagnosis? Patel and Groen (1986) found evidence, based on verbal protocols obtained during a task involving explanation of medical cases, that the overall direction of reasoning may reverse from less skilled to expert diagnosticians. The protocols of less experienced subjects tended to refer first to diseases and then to the symptoms they might account for (i.e., following the cause-to-effect direction), whereas the protocols of experts tended to move directly from symptoms to a diagnosis (i.e., following the effect-to-cause direction).

There are at least two possible explanations of such directional shifts in protocols. One possibility is that the shift simply reflects the experts' greater speed in performing the same basic reasoning process, which for subjects at all levels of expertise may make use of links directed from causes to effects. That is, diagnostic inferences based in part on cause-to-effect relationships may simply be omitted in the protocols of experts' more fluent reasoning processes. Eddy's (1982) report of preferential use of cause-to-effect conditional probabilities by practicing physicians performing diagnostic tasks might be interpreted as support for this possibility. A second possibility is that with expertise the diagnostic task is actually restructured, with directed effect-to-cause links being learned as a consequence of practice in diagnosing cases. Clearly the role of expertise in predictive and diagnostic inference requires further investigation.

Does Conditioning Involve Causal Inference?

Given that the presented experiments clearly refute recent attempts to reduce higher order types of learning to associative learning, one can of course raise the question of whether even lower order types of learning, such as classical conditioning, are explained adequately by associative learning theories. For example, despite the impressive range of findings accounted for by the Rescorla-Wagner model, the model has a variety of well-known empirical shortcomings (Gallistel, 1990; Holyoak, Koh, & Nisbett, 1989). The present findings for human causal induction raise the question of whether animals may also differentiate between predictive and diagnostic learning. Although we are unaware of any direct evidence for this possibility, some research indicates that animals can separate the temporal order of events from that of presented cues. For example, Matzel, Held, and Miller (1988) showed that rats can form representations of temporal orderings of events, even in situations where the order of presentation of the relevant information does not conform to the temporal ordering of the events within the mental model the rats seem to acquire. In Matzel et al.'s study, rats first learned that a 5-s click immediately preceded the onset of a 5-s tone. Then, in a second learning phase, the rats learned that a 5-s shock immediately preceded a 5-s tone. Interestingly, the results of the experiment suggested that the rats were able to synthesize these two learning experiences into a unified temporal representation. As expected, the rats did not show any signs of backward conditioning: When presented with the tone cue, they did not show any fear reactions. However, they seemed to expect shock when presented with the click cue. Although the rats never directly experienced clicks being paired with shocks, the clicks could be expected to predict the shock when both learning phases are integrated. Such findings indicate that some lower animals are able to dissociate the associative level from a higher order mental-model level. Is it possible, after all, that lower-order associative learning should be reduced to higher order causal induction, rather than vice versa?

References

Alloy, L. B., & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review, 91,* 112–149.

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition, 11,* 211–227.

Billman, D. (1989). Systems of correlations in rule and category learning: Use of structured input in learning syntactic categories. *Language and Cognitive Processes, 4,* 127–155.

Carlson, R. A., & Dulany, D. E. (1988). Diagnostic reasoning with circumstantial evidence. *Cognitive Psychology, 20,* 463–492.

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

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. (1992). Covariation in natural causal induction. *Psychological Review, 99,* 365–382.

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, England: Cambridge University Press.

Dickinson, A., & Shanks, D. (1985). Animal conditioning and human causality judgment. In L. G. Nilsson & T. Archer (Eds.), *Perspectives on learning and memory* (pp. 167–191). Hillsdale, NJ: Erlbaum.

Dickinson, A., Shanks, D., & Evenden, J. (1984). Judgment of act-outcome contingency: The role of selective attribution. *Quarterly Journal of Experimental Psychology, 36A,* 29–50.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249–267). Cambridge, England: Cambridge University Press.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin, 99,* 3–19.

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.

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

Gelman, S. A. (1988). The development of induction within natural kind and artifact categories. *Cognitive Psychology, 20,* 65–95.

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

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

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery.* Cambridge, MA: MIT Press.

Holyoak, K. J., Koh, K., & Nisbett, R. E. (1989). A theory of conditioning: Inductive learning within rule-based default hierarchies. *Psychological Review, 96,* 315–340.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior* (pp. 276–296). New York: Appleton-Century-Crofts.

Keil, F. C. (1989). *Concepts, kinds, and conceptual development.* Cambridge, MA: MIT Press.

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.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review, 82,* 276–298.

Matzel, L. D., Held, F. P., & Miller, R. R. (1988). Information and expression of simultaneous and backward associations: Implications for contiguity theory. *Learning and Motivation, 19,* 317–344.

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist, 44,* 1469–1481.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology, 19,* 242–279.

Mellers, B. A. (1980). Configurality in multiple-cue probability learning. *American Journal of Psychology, 93,* 429–443.

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

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289–316.

Patel, V. L., & Groen, G. J. (1986). Knowledge based solution strategies in medical reasoning. *Cognitive Science, 10,* 91–116.

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

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

Peng, Y., & Reggia, J. A. (1990). *Abductive inference models for diagnostic problem-solving.* New York: Springer-Verlag.

Poltrock, S. E., & Foltz, G. S. (1988). APT PC and APT II: Experiment development systems for the IBM PC and Apple II. *Behavioral Research Methods, Instruments, and Computers, 20,* 201-205.

Putnam, H. (1975). The meaning of "meaning." In H. Putnam (Ed.), *Mind, language, and reality: Philosophical papers* (Vol. 2, pp. 215-271). Cambridge, England: Cambridge University Press.

Reichenbach, H. (1956). *The direction of time.* Berkeley: University of California Press.

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., 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.

Shaklee, H., & Tucker, D. (1980). A rule analysis of judgments of covariation between events. *Memory & Cognition, 8,* 208-224.

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

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

Shanks, D. R. (1990b). Connectionism and the learning of probabilistic concepts. *Quarterly Journal of Experimental Psychology, 42A,* 209-237.

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: Advances in research and theory* (Vol. 21, pp. 229-261). San Diego, CA: Academic Press.

Suppes, P. (1970). *A probabilistic theory of causality.* Amsterdam: North-Holland.

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

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* (pp. 49-72). Hillsdale, NJ: Erlbaum.

Waldmann, M. R., & Holyoak, K. J. (1990). Can causal induction be reduced to associative learning? *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 190-197). Hillsdale, NJ: Erlbaum.

Ward, W. D., & Jenkins, H. M. (1965). The display of information and the judgment of contingency. *Canadian Journal of Psychology, 19,* 231-241.

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

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

Received December 31, 1990
Revision received September 3, 1991
Accepted September 10, 1991 ■