Journal of Experimental Psychology:
Learning, Memory, and Cognition
1991, Vol. 17, No. 5, 837-854

0278-7393/91/$3.00

Trial Order Affects Cue Interaction in Contingency Judgment

Gretchen B. Chapman
University of Pennsylvania

Recent research on contingency judgment indicates that the judged predictiveness of a cue is dependent on the predictive strengths of other cues. Two classes of models correctly predict such cue interaction: associative models and statistical models. However, these models differ in their predictions about the effect of trial order on cue interaction. In five experiments reported here, college students viewed trial-by-trial data regarding several medical symptoms and a disease, judging the predictive strength of each symptom with respect to the disease. The results indicate that trial order influences the manner in which cues interact, but that neither the associative nor the statistical models can fully account for the data pattern. A possible variation of an associative account is discussed.

The ability to detect predictive relationships among environmental events grants humans and other animals a distinct benefit. Therefore, the mechanisms underlying this ability are of considerable interest. Recent research with humans on judgments of contingencies has shed light on these mechanisms. It has suggested two classes of theoretical models that capture many of the central findings.

One attractive account of contingency judgment has emerged from associative learning models. Examples of such models were first developed to explain associative learning in animals. Some investigators (e.g., Alloy & Abramson, 1979; Dickinson, Shanks, & Evenden, 1984; Gluck & Bower, 1988) noted parallels between judgments of contingency by human subjects and classical conditioning in animals. These authors argued that because both tasks involve the detection of predictive relationships, they may use a common mechanism. Models of classical conditioning have recently sparked new interest (Gluck & Bower, 1988) because of their formal similarities to adaptive network models. McClelland, Rumelhart, and the PDP Research Group (1986; Rumelhart, McClelland, & the PDP Research Group, 1986) noted that associative network models can provide accounts of a variety of human cognitive processes, including judgments of contingency.

A second class of models originates from the view that humans make intuitive applications of certain formal statistical methods to everyday problems (Peterson & Beach, 1967). According to this view, humans detect a predictive relation in the same manner as would a statistical test. In line with this view, a number of theorists developed psychologically plausible versions of statistical models (e.g., Cheng & Novik, 1990; Kelley, 1973).

Both of these classes of models have the ability to predict certain fundamental phenomena of contingency judgment. In particular, they are both able to account for the result that cues interact with one another in the prediction of other events. However, they differ in their accounts of the details of that interaction in such a way as to permit an experimental test.

Cue Interaction

Early investigators of contingency judgment (e.g., Ward & Jenkins, 1965) asked subjects to judge the contingency between a single cue and a single outcome. Descriptive models that emerged from such studies naturally envisioned the subject as evaluating a cue in isolation of alternative cues. However, recent research using multiple cues has indicated that cues interact such that the predictive strength of one cue influences the judged predictiveness of alternative cues (e.g., Chapman & Robbins, 1990; Dickinson & Shanks, 1985; Dickinson et al., 1984; Gluck & Bower, 1988; Shanks, 1985, 1986, 1989; Shanks & Dickinson, 1987; Wasserman, 1990). Two examples of such cue interaction are blocking and conditioned inhibition, both phenomena first discovered in animal conditioning experiments (e.g., Kamin, 1968, 1969; Mackintosh 1975a; Pavlov, 1927; Rescorla, 1969, 1981; Rescorla & Holland, 1977), but subsequently demonstrated in contingency judgment tasks with humans (e.g., Chapman & Robbins, 1990; Dickinson et al., 1984; Gluck & Bower, 1988; Shanks, 1985).

Chapman and Robbins (1990) provided examples of both these cue-interaction phenomena. In their experiments, subjects viewed a series of trials containing information about a fictitious stock market. Periodically they were asked to judge the extent to which the rise in price of a number of fictitious stocks was predictive of the rise in price of the entire market. In the first phase of the blocking experiment, the rise in price of one stock (P) was predictive of a rise in the market, and a

Portions of this work were submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the University of Pennsylvania. This research was supported by National Science Foundation Grant BNS 88-03514 to Robert A. Rescorla.

I thank Robert A. Rescorla for supervision of this research and J. Baron and P. Rozin for valuable comments. S. Robbins and C. Tassoni provided helpful discussions. J. Nachmias of the University of Pennsylvania Psychology Department provided funds for subject payment. D. Shanks and two anonymous reviewers made helpful comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Gretchen B. Chapman, Department of Decision Sciences, University of Pennsylvania, 1300 Steinberg Hall-Dietrich Hall, Philadelphia, Pennsylvania 19104-6366.

rise in price of another stock (N) was not. In a second phase, two types of compound trials occurred, one involving P together with a novel stock B and another involving N together with a novel stock C. During this phase, both types of compounds predicted the market's rise. Despite the fact that B and C had the same individual relationship to the rise in the market, subjects judged B as less predictive than C. The presence of the highly predictive stock P blocked B from acquiring predictive strength. The nonpredictive stock N had no such effect on stock C, which did acquire predictive strength. Thus, stock B is identified as a blocked cue, whereas stock C served as a comparison cue. This experiment constitutes one demonstration of cue interaction. Similar results have been obtained by other investigators (Gluck & Bower, 1988; Shanks, 1985, 1989).

In the conditioned-inhibition experiment conducted by Chapman and Robbins (1990), the rise in price of a single stock (P) was followed by a rise in the market price. On other trials, the price of stock P rose in conjunction with the price of another stock (I), but that compound was followed by no rise in the market. Stock I was rated as more negative than a comparison stock (N) whose price rose in isolation and was followed by no rise in the market price. Although stocks I and N had the same individual relationship with the outcome, the presence of the highly predictive stock P caused stock I to acquire negative predictive strength, also called inhibition. This conditioned inhibition experiment is another demonstration of cue interaction.

Results such as blocking and conditioned inhibition demonstrate that the predictive strength of one cue is importantly affected by the predictive strengths of other copresent cues. Any adequate account of contingency judgment in humans must be able to characterize such interaction. Two classes of models endeavor to do so.

Associative Models

Within the field of animal learning, several associative models were developed to explain the blocking and conditioned inhibition findings (Mackintosh, 1975b; Pearce & Hall, 1980; Rescorla & Wagner, 1972). One such conditioning model, proposed by Rescorla and Wagner (1972), interprets blocking and inhibition as parallel phenomena and shares some features with the network and statistical models discussed later. According to the Rescorla-Wagner model, the predictive strength of each cue is updated on each trial in which that cue is presented. The amount that the predictive strength of a given cue changes on a particular trial is a direct function of the discrepancy between the outcome that actually occurs on that trial and the outcome that was expected for that trial on the basis of all cues present. Consequently, predictive strengths are adjusted in such a way as to reduce this discrepancy.

The Rescorla-Wagner model explains cue interaction because the discrepancy depends on the predictive strengths of all cues present on a trial. In the blocking experiment described previously, stock P first gained a large amount of predictive strength, whereas N gained little if any. Later P and B were presented in compound and followed by the outcome. On these trials, the outcome was fully predicted on the basis of P, and thus there was no discrepancy between the actual and expected outcomes. Consequently, no strengths were adjusted and B's strength remained low. On trials where N was presented in compound with C, a large discrepancy existed between the actual and expected outcomes. As a result, the strengths of both symptoms N and C increased.

The Rescorla-Wagner model explains conditioned inhibition in a parallel manner. In the Chapman and Robbins (1990) experiment, P gained positive predictive strength on trials in which it was presented and the market rose. On the first few trials in which P and I were presented in compound, the expected outcome was quite high because P was quite predictive and stock I had as yet no strength. Because the actual outcome was zero, there was a negative discrepancy between the actual and expected outcomes. As a result, the strengths of both cues were adjusted down. Because the starting strength of I was zero, a decrease in strength resulted in the acquisition of negative predictive strength. In contrast, N was always presented alone; thus, there was no discrepancy between actual and expected outcomes, and N's strength remained at zero.

Other associative models of conditioning (e.g., Mackintosh, 1975b; Pearce & Hall, 1980) differ in mechanism, but share the prediction of cue-interaction results. They also share with the Rescorla-Wagner model the property that cue strengths are updated on a trial-by-trial basis.

Gluck and Bower (1988) recently pointed out the similarities between associative models of animals learning and adaptive network models (McClelland et al., 1986; Rumelhart et al., 1986). In fact, the Rescorla-Wagner model can be viewed as one version of an adaptive network model.

Adaptive networks consist of many processing units or nodes connected by weighted, unidirectional links of activation. The nodes are separated into layers: an input layer, an output layer, and a number of hidden unit layers. The activation of a given node is determined by the weighted inputs from all the nodes in the adjacent layer. When a stimulus is presented to the network, a set of input nodes is activated. These nodes activate nodes in the next layer. The resulting pattern of activation in the output layer corresponds to some response of the system or to the estimated outcome. The network then receives some feedback as to the desired output pattern. This feedback corresponds to the actual outcome. The weights of the links are then adjusted to bring the output pattern closer to the desired pattern. These weights are analogous to the predictive strengths of the input nodes.

Gluck and Bower (1988) described a network model that is formally equivalent to the Rescorla-Wagner model and explored its suitability as a model of contingency judgment. This network had an input layer and an output layer but no hidden layers. The weights of the links from input nodes to output nodes were altered according to the least mean squares (LMS) rule, which Gluck and Bower (1988) noted is equivalent to the Rescorla-Wagner learning rule, because according to both rules the change in strength of a particular node or cue is some fraction of the discrepancy between actual and expected outcomes. Strengths are updated in a trial-by-trial

manner, and the strength of a given cue is updated only on trials on which that cue is present.

The class of adaptive network models encompasses a wide variety of models (McClelland et al., 1986; Rumelhart et al., 1986). Consequently, particular network models often differ from the Gluck and Bower (1988) model. For example, some network models use a slightly different learning rule or include layers of hidden nodes, which increase the scope of problems to which the network can be applied. Additional features of other network models will be explored in the General Discussion section. For purposes of comparison with the statistical models discussed next, conditioning models and network models are referred to as associative models.

Statistical Models

Associative models have been developed strictly as models to describe and explain how people judge contingencies; they do not prescribe how people ought to assess predictive relations. As such, they can be contrasted with normative models. The normative or optimal method for computing the relationship between two events is to perform a statistical test. A sizable body of research has viewed people as intuitive statisticians (e.g., Nisbett, Krantz, Jepson, & Kunda, 1983; Peterson & Beach, 1967) capable of mentally performing computations that are roughly equivalent to statistical tests. In judging the relations between multiple predictive cues and a single outcome, a reasonable statistical technique to use is a multiple linear regression. One might view subjects as computing rough equivalents of regression weights for each of the predictive cues under consideration.

The standard use of a multiple linear regression model is to predict the value of a criterion (or outcome) based on a number of predictors or cues. Weights (or predictive strengths) are assigned to the cues to minimize the squared error. The error is defined as the difference between the expected outcome and the actual outcome, and the squared error is computed by squaring this error and summing over all trials. Interestingly, such a process is similar to associative models in that selection of appropriate predictive strengths acts to minimize some measure of the discrepancy between the actual outcome and the expected outcome.

Computation of linear regression weights would result in the cue-interaction effects discussed previously. The weights that minimize error in the blocking experiment are a weight of 1 for stock P and a weight of 0 for stock B. These weights ensure that the outcome will be accurately predicted on P trials but not overpredicted on PB trials. In contrast, N will be assigned a weight of 0 and C a weight of 1. These weights guarantee that the outcome will be predicted on NC trials but not on N trials. Thus, a regression analysis assigns stock B a lower weight than it assigns stock C. In the inhibition experiment, P would be assigned a weight of 1, I a weight of -1, and N a weight of 0. With these weights, the outcome would be predicted on P trials but not on PI or N trials. Thus, cue I is given a weight lower than that given cue N.

Although some authors (e.g., Koh & Meyer, 1991) suggested that in judging contingent relations people perform an operation akin to statistical regression, others (e.g., Cheng & Novick, 1990) viewed normative statistical models as implausible psychological models because they are quite computationally complex. Consequently, several psychological models have been proposed that maintain many important features of statistical models but are more computationally tractable.

Extensive work on contingency judgment (e.g., Beyth-Marom, 1982; Crocker, 1981; Shaklee & Tucker, 1980; Ward & Jenkins, 1965; Wasserman, Chatlosh, & Neunaber, 1983) proposed that, in judging contingencies, people store frequencies of different types of trials in a 2×2 table, which crosses the presence or absence of a cue with the presence or absence of the outcome. At least on some occasions, people then compute the normative Δp statistic, which involves comparing the probability of the outcome given the presence of the cue with the probability of the outcome given the absence of the cue. This computation yields a judgment formally similar to the simple correlation between the cue and outcome; however, it avoids the complex computations required by a statistical regression. Use of the Δp rule, however, cannot account for cue-interaction results (see Chapman & Robbins, 1990), and therefore cannot provide a complete account of contingency judgment.

Cheng and Novik (1990) proposed a probabilistic contrast model of causal judgment, which can be viewed as an extension of the Δp statistic to the case of multiple cues. This model proposes that subjects compute contrasts between the proportion of times the outcome occurs with a particular value on a dimension versus other values on that dimension. Subjects use these contrasts to infer the predictive strength of multiple causal factors, as well as interactions among those factors. The probabilistic contrast model is based on Kelley's (1973) earlier analysis of variance model. Both of these models propose that contingency judgment involves computing a simplified version of a normative statistical test.

These simplified statistical models are similar to normative statistical techniques in two important ways. First, each can explain cue interaction in the same manner as multiple linear regression. Second, their predictions are independent of the order in which information is presented. For purposes of comparison with associative models, all of these normative statistical and simplified statistical models are referred to as statistical models.

Comparisons Between Cue-Interaction Models

The two classes of models discussed are similar in that they calculate predictive strengths that minimize some measure of the discrepancy between actual and expected outcomes. As Gluck and Bower (1988) discussed, the Rescorla-Wagner learning rule results in the selection of weights that minimize the total squared error; thus, an associative model will approach the linear regression solution (Stone, 1986). Therefore, associative and statistical models are quite similar in their accounts of cue interaction.

Although both types of models do have a common goal of minimizing the discrepancy between actual and expected outcomes, they do not always arrive at the same solution. These differences are a result of the method each uses to reduce the discrepancy. When applying an associative model,

such as the Rescorla-Wagner model, the predictive strength of each of the cues is updated after each trial in which that cue appears. The strength of a given cue is altered only on trials in which that cue is present, and each trial has an effect on the predictive strengths only once. Thus, the order in which the trials are presented may affect the resulting predictive strengths.

By contrast, a statistical model does not update strengths after each trial; instead, the information from each trial is stored for later use. When the subject wishes to evaluate the cues, a statistical test is computed based on the summarized data. The order in which the trials were presented is not stored and therefore cannot have any effect on the weights selected. A particular trial has the same influence regardless of when in the trial sequence it was presented.

Associative models anticipate that trial order will affect both of the cue-interaction phenomena discussed previously. Consider a blocking study in which cue A is first presented in a positive contingency with the outcome (A+). Later cue A and cue B appear together and jointly predict the outcome (AB+). Cue B will have a low associative strength because when it occurs in the second phase there is a small discrepancy between the actual outcome and the outcome expected on the basis of the presence of A and B. Alternatively, if this trial order were reversed, such that the AB+ trials preceded the A+ trials, blocking should not occur. In the first phase, there would be a large discrepancy between the actual outcome and the outcome expected on the basis of A and B; thus, both A and B would gain some strength. The subsequent phase in which A alone predicts the outcome would not act to lower B's strength because, according to the Rescorla-Wagner model, the strength of a cue cannot be altered if that cue is not present. In other words, subjects do not retrospect and reevaluate old conclusions about one cue based on new evidence about a different cue.

In contrast, statistical models predict that both trial orders should yield the same predictive strengths. A linear regression model acts to minimize error over all trials. The weights that minimize error in the previous example are a weight of 1 for A and a weight of 0 for B regardless of the trial order.

Associative models make similar predictions about order effects in other cue-interaction phenomena. Conditioned inhibition, for example, should be more effective if the P+ trials are presented before the PI− trials rather than the reverse order. Statistical models predict that such trial order should have no effect.

Other models of contingency judgment can be usefully compared with the two classes discussed here on the basis of their predictions about the importance of trial order. For example, Dickinson and Shanks (1985) proposed a theory that predicts that the two blocking trial orders discussed previously should be equally effective in producing cue interaction. Although their theory is similar to associative models in important ways, for the present analysis it would be grouped with the statistical models because it predicts no effect of trial order. Providing another example, Holyoak, Koh, and Nisbett (1990) proposed a rule-based theory of conditioning, which could also be applied to contingency judgment. Although this theory is quite different from the associative models discussed here, it predicts that cue interaction will be affected by trial order (Holyoak et al., 1990, p. 327). Therefore, the present analysis groups this rule-based theory with the associative models.

The present experiments explored these differing predictions about the role of trial order in blocking and conditioned inhibition to distinguish different classes of models of contingency judgment. Experiment 1 provided a demonstration of blocking. Experiments 2 and 3 explored the effect of trial order on blocking. Experiments 4 and 5 examined trial order effects in conditioned inhibition.

Experiment 1

The purpose of Experiment 1 was to demonstrate blocking with a contingency judgment task. This demonstration was a necessary prerequisite to an investigation of the role of trial order in cue-interaction phenomena.

In this experiment, the subjects' task was to use information about symptoms exhibited by fictitious medical patients to predict the presence of a fictitious disease. Information about the symptoms and disease was given in a series of trials framed as different patient histories. Similar medical diagnosis tasks have been used in other investigations of contingency judgment (e.g., Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Gluck & Bower, 1988; Medin & Edelson, 1988).

The trials were divided into two phases (see Table 1). In Phase 1 symptom P was a good predictor of the disease, whereas symptom N was nonpredictive of the disease. A third symptom (M) was not displayed in this phase. In Phase 2, pairs of symptoms were displayed together and always present with the disease. On some trials P and B were displayed together, whereas on other trials M and C_M were displayed together. On still other trials N and C_N were displayed together. Periodically subjects were asked to assess the predictive strength of each symptom.

If subjects evaluate predictive cues in isolation, then B, C_M, and C_N should be rated similarly because all bear the same individual relation to the disease. By contrast, if copresent cues compete for predictive strength, then the high predictive strength of P should result in a low strength for B (the putatively blocked stimulus) relative to symptoms C_M and C_N, which serve as two comparison stimuli against which to evaluate the cue-competition effect. The reasons for using two comparison stimuli are elaborated in the *Results and Discussion* section.

Table 1
Design for Experiment 1

Phase 1	Phase 2
P → disease	PB → disease
	MC_M → disease
N → no disease	NC_N → disease

Note. On completion of each phase, subjects were asked to rate each of the six symptoms.

Method

Subjects and apparatus. The subjects were 24 undergraduates from Philadelphia-area colleges. They were paid $5 per hour for this experiment (which lasted approximately 30 min) and a number of other unrelated experiments. Subjects were run individually using IBM XT computers with color monitors.

Procedure. When a subject first sat down at a computer, instructions appeared explaining that she or he would be asked to examine information about a number of fictitious medical patients. It was explained that each patient would exhibit a set of symptoms. The subject's task was to predict how likely it was that each patient suffered from the fictitious disease called morolis. The subject was further instructed that periodically she or he would be asked to make a prediction about a diagnosis based on information about only one symptom.

After the instructions, the first patient history appeared on the screen. At the top of the screen was displayed a patient number that progressed with each trial. A number of symptoms were listed in alphabetical order. The symptom names that could appear were asthma, bruises, coughing, dizziness, earache, and fever. Below the names of the symptoms was printed the question "What symptoms are present (A, B, C, D, E, F)? Type first letter of symptom or press space bar." This response was required to ensure that the subject encoded each of the displayed symptoms. After the subject had typed the first letter of each displayed symptom or pressed the space bar if no symptoms were present, everything but the list of symptoms was erased. The subject was then asked "How probable is it that the patient has morolis (0–100)?" Ratings were requested to encourage active engagement in the task. After the subject typed a number between 0 and 100, the part of the screen below the symptom list was erased. The subject then received feedback; the screen displayed either the sentence "The patient does have morolis," with the word *morolis* in flashing green letters, or the sentence "The patient does not have morolis," with the word *not* in red letters. The subject then pressed a key to advance to the next trial.

At the end of each of the two phases, the subject was asked to rate the six symptoms. At such a point, the subject was instructed to imagine for each symptom that she or he received information that a patient displayed that symptom, but that nothing was known about any other symptoms. The subject was asked to use all of the information received up to that point to estimate the likelihood that such a patient suffered from morolis. This rating was expressed as a number between 0 and 100. The subject was allowed to correct and adjust the six ratings with the stipulation that no two symptoms could have the same rating. Rating ties were not allowed so that subjects would be forced to rank the blocked symptom relative to the two comparison stimuli. At the end of the experiment, subjects were thanked for their participation and were paid.

Each subject was randomly assigned to one of six conditions of a Latin square design to counterbalance which symptom named played each stimulus role. The different roles played by the symptoms are indicated as P, M, N, B, C_M, and C_N. This experiment contained two phases: element training and compound training.

The element-training phase contained 36 patients or trials. On 12 of these trials, only symptom P was displayed and morolis was present (abbreviated as P+ trials). There were also 12 trials in which only symptom N was displayed and morolis was absent (N−). On the other 12 trials, no symptoms were present and morolis was absent. These trial types were presented in a quasirandom order in the first phase, and after the 36th trial subjects were asked to rate the symptoms.

The compound-training phase also consisted of 36 trials; subjects received 9 each of 4 trial types given in intermixed order. On 3 of these trial types (PB+, MC_M+, NC_N+), the indicated compound was displayed and the disease was present. On the remaining trials, no symptoms were present and the disease was absent. At the end of this phase, subjects were asked to rate the symptoms.

Results and Discussion

Statistical comparisons were made using nonparametric statistics to minimize the assumptions made about the relationship between the perceived predictive strength of a symptom and the rating given to express that strength. Within-subject pairwise comparisons were made using the two-tailed Wilcoxon T test (Siegel, 1956). For each statistic, the degrees of freedom represent the number of subjects involved in the comparison. A number less than the number of subjects in the experiment indicates that some of the subjects exhibited ties. The value of the statistic is a measure of the number and magnitude of differences in the direction opposite to that expected; thus, smaller values are more likely to be statistically reliable. The number of subjects who exhibited differences in the predicted direction and in the opposite direction and the number who gave ties are indicated in parentheses after each statistic. For example, (22, 2, 0) indicates that 22 subjects rated the symptoms in the predicted direction, 2 rated them in the opposite direction, and no subjects gave ties.

Figure 1 shows the mean ratings given as a function of the symptom rated. The top panel shows the rating given after the first phase. Symptom P was rated higher than M and N, $T(24) \leq 5$, $p < .01$ (22, 2, 0), whereas M was rated higher than N, $T(24) = 10$, $p < .01$ (21, 3, 0). It was expected that P be rated higher than M and N because P had been presented as a positive predictor of the disease whereas M and N had not. This difference is important because it enables P to compete successfully with B for predictive strength in the second phase. The difference between M and N may be attributed to M's gaining some generalized strength from P. Presentations of N may have protected it from such generalization. Alternatively, M may have maintained a neutral strength, whereas N acquired some negative predictive strength.

At the end of the first phase, ratings of B, C_M, and C_N did not differ from one another, $T(24) \geq 95.5$, $p \geq .23$ (15, 9, 0). Because none of these symptoms had thus far been displayed, it was expected that they would be rated similarly. It is interesting to note that B, C_M, and C_N, along with symptom M, were assigned positive ratings of about 30 despite the fact that they had not been presented previously. These positive ratings suggest several possible explanations. Subjects may have based their ratings of these stimuli on the base rate of morolis in the previous block of trials. Another possibility is that, given no information, subjects assume that any symptom is slightly predictive of the disease. Alternatively, as discussed previously, these symptoms may have gained some generalized strength from symptom P.

The bottom panel of Figure 1 presents ratings given after the second phase. Here P continued to be rated higher than M and N, $T(24) \leq 17.4$, $p < .01$ (22, 2, 0); M was also rated higher than N, $T(24) = 14$, $p < .01$ (20, 4, 0).

The data of primary interest are the ratings of stimuli B, C_M, and C_N given after Phase 2. After compound training, B

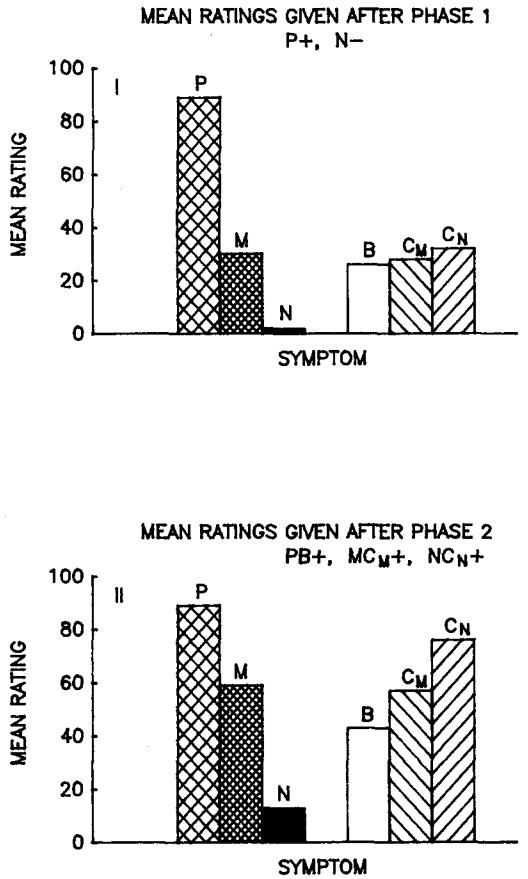

Figure 1. Mean ratings given to each symptom after each of the two phases of Experiment 1. (In Phase 1, symptom P was presented as a good predictor of the disease, whereas symptom N was not predictive. In Phase 2, P was presented in compound with B, M was presented in compound with C_M, and N was presented in compound with C_N. The disease was present on each compound trial.)

was rated lower than both C_M, $T(24) = 61$, $p < .01$ (17, 7 0), and C_N, $T(24) = 48.5$, $p < .01$ (17, 7, 0). Symptom C_M was also rated less than C_N, $T(24) = 64$, $p < .02$ (16, 8, 0). The low rating of B relative to the two comparison stimuli constitutes the blocking effect. Symptoms B, C_M, and C_N were embedded in identical individual relationships to the disease, but they received different ratings depending on the differing predictive strengths of their partner symptoms (P, M, and N).

The lower rating of C_M relative to C_N is understandable given the different associative strengths of M and N. After the first phase, symptom M evidenced more predictive strength than N; thus, it may be inferred that M was better able to compete with C_M than was N with C_N.

Symptoms C_M and C_N both serve as comparison stimuli against which to evaluate the blocking effect, although it is unclear which provides the more appropriate comparison. Symptom C_N offers the advantage that, like B, it was presented in compound during Phase 2 with a preexposed stimulus; it differed from B only in the prior treatment of its copresent stimulus (N). Both associative and statistical models predict that, given a sufficient number of trials, presentation of symptom N in the first phase should have resulted in a zero predictive strength for N, making it nonpredictive of the disease. It is possible, however, that contrary to these models this preexposure of N could have resulted in the accrual of negative predictive strength by N. This negative strength would then have facilitated the acquisition of positive predictive strength to C_N, resulting in an exaggerated depiction of the blocking result. It is important to note that although the difference between the ratings of B and C_N may represent more than the blocking effect, the difference is unequivocally the result of cue interaction.

C_M provides an alternative comparison for B. It is doubtful that the nonpresented symptom M could have acquired negative predictive strength in Phase 1; thus, its partner stimulus, C_M, would have no additional advantage in gaining positive predictive strength. As postulated earlier, however, symptom M may have received some generalized positive strength during the first phase, as suggested by the ratings given after Phase 1. If this were the case, M would have been capable of causing C_M to be partially blocked. Thus, use of C_M as a comparison stimulus may result in an underestimation of the blocking result.

Regardless of the comparison stimulus chosen, however, Experiment 1 demonstrates the phenomenon of blocking. In Experiments 2 and 3, a blocked stimulus is compared with a stimulus treated as C_N has been treated here. This comparison is taken as the assessment of blocking with the understanding that, according to some interpretations, this comparison yields an overestimate of blocking. Put another way, the differences revealed through this comparison could be due to the treatment of P, the treatment of N, or a combination of both treatments. Nevertheless, such differences are undeniably the result of cue interaction, and thus the more general goal of examining order effects in cue interaction may be met.

Experiment 2

The first experiment provided a demonstration of what might be termed *forward blocking.* In a forward-blocking procedure, one stimulus (symptom P) is established as a good predictor before compound presentations with a second stimulus (symptom B). Conversely, in a backward-blocking procedure, the trials are presented in the opposite order: A stimulus is established as a singly good predictor only after compound presentations with another stimulus.

The second experiment served to compare directly forward- and backward-blocking effects. As discussed early here, statistical models predict that blocking will be unaffected by trial order; consequently, forward- and backward-blocking procedures should be equally effective in attenuating the acquisition of predictive strength by the putatively blocked stimuli. Conversely, associative models predict that blocking will be effective only in the forward direction. Thus, a stimulus in a forward-blocking procedure should gain less strength than a stimulus in a backward-blocking design.

Experiment 2 consisted of three phases, as presented in Table 2. In the first phase, one symptom (P) was presented as a good predictor of the disease, whereas another symptom (N) was presented as nonpredictive. In the second phase, two

Table 2
Design for Experiment 2

Phase 1	Phase 2	Phase 3
P → disease	PF → disease	P → no disease
N → no disease	NB → disease	N → disease

Note. On completion of each of the three phases, subjects were asked to rate each of the four symptoms.

pairs of symptoms were presented (PF and NB) and were always followed by the disease. In the third phase, the treatments of P and N were reversed: P was presented as not predictive of the disease, whereas N was presented as a positive predictor. After the third phase, symptoms were presented individually in test trials, which provided an additional opportunity to measure the judged predictive strengths of the symptoms.

Phases 1 and 2 constitute a forward-blocking design for F similar to that used in Experiment 1. Symptom P was established as a good predictor before the PF compound trials, making F a forward-blocked stimulus. By contrast, Phases 2 and 3 constitute a backward-blocking design for symptom B. Symptom N was established as a good predictor only after the NB compound trials, making B a backward-blocked stimulus. Symptoms F and B were treated identically in all respects except the time at which the partner symptom for each was established as a good predictor of the disease. This design allows for a within-subject comparison of the magnitudes of forward and backward blocking. If the blocking phenomenon is unaffected by trial order, F and B should be similarly rated after Phase 3. By contrast, if forward blocking is more effective than backward blocking, then F should be rated lower than B.

Method

Subjects and apparatus. The subjects were 24 undergraduates from area colleges. They were paid $5 per hour for this experiment and a number of other unrelated experiments. The same computers and program were used as in Experiment 1 with the exception of the particular trial types administered.

Procedure. The instructions given were the same as those in Experiment 1, as were the details of trial presentation and elicitation of ratings. Each subject was assigned to one of four conditions of a Latin square to counterbalance which symptom name corresponded to each stimulus role. Six subjects were assigned to each of the four counterbalance conditions. The symptom names used were asthma, bruises, coughing, and dizziness.

Experiment 2 was divided into three phases. Phases consisted of 12 each of 3 intermixed trial types. During Phase 1, on 12 trials symptom P was displayed and the disease was present (P+). On 12 other trials, symptom N was displayed and the disease was absent (N−). On the remaining 12 trials, no symptoms were displayed and the disease was absent. The second phase contained 12 trials in which both symptom P and symptom F were displayed together and the disease was present (PF+). There were also 12 trials on which both symptoms N and B were present and the disease was present (NB+). On the remaining 12 trials in this phase, no symptoms were displayed and the disease was absent. In the third phase, there were 12 trials on which symptom P was displayed and the disease was absent (P−). On 12 other trials, symptom N was present and the disease was present (N+). On 12 additional trials, no symptoms were present and the disease was absent. After each of the three phases, subjects were asked to rate each symptom.

Test trials followed the third phase; each symptom was presented individually on each of three trials and the disease was never present. Likelihoods given on these trials served as an additional measure of the predictive strengths of the symptoms. After these test trials, subjects were queried about their memory for which symptoms had previously appeared in compound together.

Results and Discussion

Figure 2 shows the primary results for Experiment 2. The first panel depicts mean ratings given after the first phase. As expected, P was rated higher than N at the end of the first phase, $T(24) = 0$, $p < .01$ (24, 0, 0). Also as expected, the ratings of the previously nonpresented symptoms F and B did not differ from one another, $T(24) = 148$, $p > .96$ (15, 9, 0). Symptoms F and B were given positive ratings near 30; as in Experiment 1, novel symptoms were rated as moderately predictive of the disease.

The ratings given at the end of the second phase are illustrated in the second panel of Figure 2. Although compounds containing both P and N had been paired with the disease, P continued to be rated higher than N, $T(24) = 3$, $p < .01$ (22, 2, 0). Of more interest, at the end of the second phase, F was rated less than B, $T(24) = 50.5$, $p < .01$ (17, 7, 0). This comparison between F and B illustrates the forward-blocking effect and is similar to that seen in Experiment 1.

The data of primary interest are the ratings given after the third phase, shown in Panel 3. The mean ratings of P and N have reversed because of the reversal of their treatments. Symptom N now appeared slightly higher than P, although this difference failed to reach significance, $T(24) = 87$, $p < .08$ (12, 12, 0). Most important, after the third phase F continued to be rated lower than B, $T(24) = 37$, $p < .01$ (18, 6, 0). This difference indicates that the forward-blocking procedure was more effective in attenuating the acquisition of predictive strength by symptom F than the backward-blocking procedure was in reducing symptom B's predictive strength. Because symptoms F and B were treated identically with the exception of the time at which the partner stimulus for each was established as a good predictor, this difference in ratings indicates that the blocking effect is influenced by trial order.

It is interesting to note that the ratings of B appear somewhat lower in Phase 3 than in Phase 2. This difference, which is reliable, $T(18) = 9.5$, $p < .01$ (16, 2, 6), could be viewed as evidence of backward blocking; as a result of the events presented in Phase 3, the ratings of B were reduced. An alternative explanation, however, is that subjects displayed a general tendency to use lower ratings after the third phase. In support of this alternative, the Phase 3 ratings of F were also lower than the Phase 2 ratings of F, $T(17) = 15$, $p < .05$ (13, 4, 7). The change in ratings of B from Phase 2 to Phase 3 is not greater than the change in ratings of F, $T(21) = 82.5$, $p > .25$ (14, 7, 3). Thus, it appears that subjects demonstrated a general lowering of ratings from the second to the third phase.

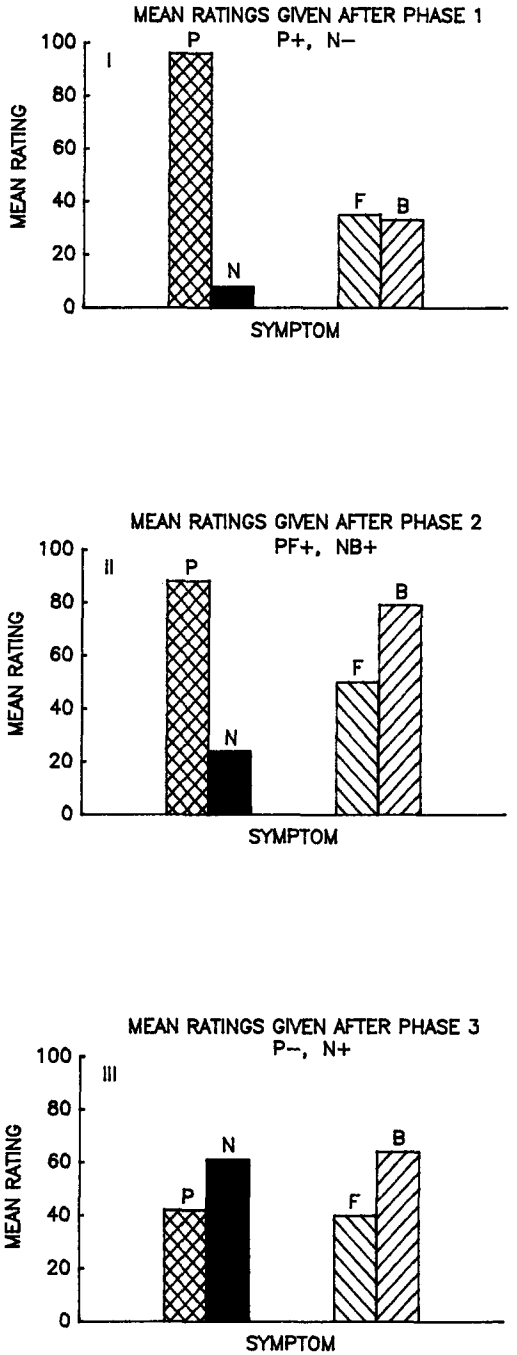

Figure 2. Mean ratings given to each symptom after each of three phases of Experiment 2. (In Phase 1, symptom P was presented as a good predictor of the disease, whereas symptom N was not predictive. In Phase 2, P was presented in compound with F, and N was presented in compound with B; the disease was present on each compound trial. In Phase 3, symptom P was presented as nonpredictive, whereas symptom N was presented as predictive of the disease.)

In this and other experiments, ratings given on the test trials were similar in all important respects to the ratings given after Phase 3. These results are therefore not presented.[1,2]

The fact that F was rated lower than B after the third phase of Experiment 2 supports the conclusion that forward blocking is stronger than backward blocking. Symptoms F and B were treated identically with the exception of the time at which the partner stimulus for each was established as a good predictor. If forward-blocking and backward-blocking manipulations were equally powerful, F and B would have been rated similarly at the end of Phase 3. The fact that they were not constitutes evidence for the importance of trial order in determining the blocking effect.

Whereas the Rescorla-Wagner model correctly predicts this effect of trial order on blocking, statistical models predict no effect of trial order. However, a statistical model, such as linear regression, might be altered to account for order effects by making an additional assumption about a primacy effect. This assumption would involve postulating that subjects processed trials more successfully at the beginning of the experiment than at the end of the experiment. This primacy would have the same effect as presenting more trials in Phase 1 than in Phase 3. It would cause P to have a higher overall predictive strength than N; consequently, even a regression model would predict that P would block F more than N would block B.

This primacy explanation can be evaluated by examining the predictive strengths of P and N. According to the primacy explanation, P should be rated higher than N at the end of the experiment. This prediction results from the assumption that more trials were processed from Phase 1 than from Phase 3. The results of Experiment 2 indicate, however, that P was not rated higher than N after Phase 3. In fact, P was actually rated lower than N in the test trials. Thus, there is no evidence that the judged difference between F and B was due to a primacy effect.

---

[1] In this and other experiments, not all subjects demonstrated the effect of interest. For example, in Experiment 2, not all subjects rated symptom F lower than symptom B. Subjects did not appear to fall into well-defined subsets, however; instead, ratings varied continuously over much of the possible range. Attempts to explain this variation across subjects were unsuccessful. In each of the five experiments, assignment to counterbalance condition had no effect on the comparison of primary interest (Kruskal-Wallis statistic ≤ 8.3, $ps > .14$, $df = 5$ in Experiment 1, $df = 3$ in Experiments 2–5). Memory for which symptoms had appeared together in compound was not consistently correlated with effect size (for Experiment 3, Spearman $r = .3$, $p < .05$; for all other experiments, $rs < .35$, $ps > .1$). Ratings given at the end of the third phase were not correlated with ratings given on test trials. In Experiments 2, 3, 4, and 5, these correlations were as follows: $r = .22$, $p > .2$; $r = .22$, $p > 1$; $r = .05$, $p > .6$; $r = .08$, $p > .5$, respectively. In Experiments 1 and 2, the difference between the ratings of the partner symptoms (e.g., P and N) was correlated with the difference between the two symptoms of interest (e.g., F and B in Experiment 2); $rs > .45$, $ps < .05$). This correlation did not hold in Experiments 3, 4, and 5, however, ($rs < .2$, $ps > .4$).

[2] A replication of Experiment 2 was conducted using probabilistic rather than deterministic relationships between the symptoms and the disease. The design was identical to that of Experiment 2 except that only 10 of every 12 disease trials actually contained the disease, and 2 of every 12 no disease trials had the disease present. The results of this replication reproduced the results of Experiment 2 in all important respects. Of primary importance, after the third phase, the mean rating of symptom F was 29, whereas that of symptom B was 44. This difference was statistically reliable, $T(24) = 77$, $p < .05$ (15, 9, 0).

It is important to note that a statistical model predicts that the ratings of F and B will always be an inverse function of the ratings of P and N. Thus, if N is rated higher than P (because of a recency effect), then N's partner, B, should be rated lower than P's partner, F. This prediction follows because if N provides a lot of information about the occurrence of the disease, then there is very little additional information value to be attributed to B; therefore, B should be assigned a low rating. Similarly, if P provides relatively little information about the disease, there is a sizable amount of variation to be explained by F; therefore, F should be assigned a high rating.

The data from Experiment 2 illustrate that the prediction of an inverse relation is not supported. In the final ratings, despite the fact that N was rated slightly higher than P, N's partner, B, was not rated lower than P's partner, F; rather B was rated higher than F. However, this direct relationship between the ratings of the blockers (P and N) and the strengths of the putatively blocking stimuli (F and B) can be explained by associative models such as the Rescorla-Wagner model. Because these models predict no retrospection, the strengths of F and B are determined not by the terminal strengths of P and N, but rather by the strengths of P and N during the second phase, when P and N co-occurred with F and B. At that time P was more predictive than N, as indexed by ratings given before and after Phase 2; consequently, F was more blocked than B. Put differently, the associative models predict that the strengths of the putatively blocked stimuli (F and B) will be an inverse function of the strengths that the blockers (P and N) possess during the compound phase. The strengths of P and N at the end of the experiment are of no consequence for the strengths of F and B. It is therefore apparent that associative models are quite capable of explaining the results of Experiment 2; conversely, these results are inconsistent with predictions made by statistical models.

Experiment 3

Experiment 2 indicates that a forward-blocking procedure results in more cue competition than a backward-blocking procedure; however, this experiment provides a relatively weak test of associative models. These associative models predict not only that the forward-trial order will be more effective than the backward-trial order in producing blocking, but also that a backward-blocking procedure will result in no cue competition.

In disagreement with this prediction, a recent experiment by Shanks (1985) found evidence for backward blocking in contingency judgment. Shanks' subjects engaged in a video game in which a mine field and the firing of a shell were each potential predictors of the explosion of a tank. In the backward-blocking condition, subjects first had the opportunity to fire shells at the tank while it traversed the mine field (AB+). In a later phase, they did not fire any shells but observed the frequency with which the tank exploded as a result of the mines (A+). Ratings of the effectiveness of the shells (B) were lower in this condition relative to a control condition that did not include the A+ observation period. From these results Shanks concluded that associative mechanisms like that of the Rescorla-Wagner model do not underlie judgments of contingency. Instead, Shanks suggested, humans engage in retrospective processing, adjusting the rated effectiveness of the shells in light of new information about the mine field.

Experiment 3 was intended to examine whether a backward-blocking procedure would produce any cue competition in the procedure used here. It therefore served a purpose similar to that of experiments by Shanks (1985). The design of Experiment 3 is illustrated in Table 3. Two pairs of symptoms were presented (PB and NC) and always followed by the disease. In a subsequent phase, P was presented as a good predictor, and N was presented as nonpredictive. Thus, symptom B was a putatively backward-blocked stimulus because its partner symptom, P, was established as a positive predictor of the disease after the compound presentations. Symptom C was a comparison stimulus because its partner symptom, N, was not established as a positive predictor.

To make this design more like that of Experiment 2, an initial phase preceded the compound presentations. This first phase involved two other symptoms. Earache (designated as E) was presented as a positive predictor, and fever (designated as V) was presented as nonpredictive.

In this experiment, symptom B was embedded in a backward-blocking design. If such a design results in cue competition, B should be rated lower than C. Statistical models predict such a result. If, as associative models predict, a backward-blocking procedure is ineffective, symptoms B and C should be rated similarly.

Method

Subjects and apparatus. The subjects were 61 undergraduates from area colleges. They were paid $5 per hour for this experiment and a number of other unrelated experiments. The same computers and program were used as in the previous two experiments, with the exception of the particular trial types.

Procedure. The instructions given were the same as those in Experiment 2, as were the details of trial presentation and elicitation of ratings. Each subject was assigned to one of four conditions of a Latin square to counterbalance which symptom name corresponded to the stimulus roles designated as C, B, P, and N. The symptom names were the same as in Experiment 1.

Experiment 3 was divided into three phases; each of the phases contained 36 trials. The first phase included 12 trials in which one symptom, earache, was presented and the disease was present (E+). In 12 other trials, fever was presented and the disease was absent (V−). There were also 12 trials in which no symptoms or disease was present. The symptom names earache and fever were used in this phase for all counterbalance conditions.

Phase 2 contained two types of compound trials. On 12 trials, symptoms N and C were presented and the disease was present (NC+). There were also 12 trials in which symptoms P and B were present and the disease was present (PB+). On the remaining 12

Table 3
Design for Experiment 3

Phase 1	Phase 2	Phase 3
E → disease	PB → disease	P → disease
V → no disease	NC → disease	N → no disease

Note. On completion of each of the three phases, subjects were asked to rate each of the six symptoms.

trials, no symptoms were displayed and the disease was absent. In the third phase, P was presented on 12 trials in which the disease was present (P+). Symptom N was displayed on 12 trials in which the disease was absent (N−). Again 12 trials occurred on which no symptom or disease was displayed. After each of the three phases, subjects were asked to rate each symptom. After Phase 3, test trials were presented, and subjects were queried about their memory for which symptoms had previously appeared in compound together.

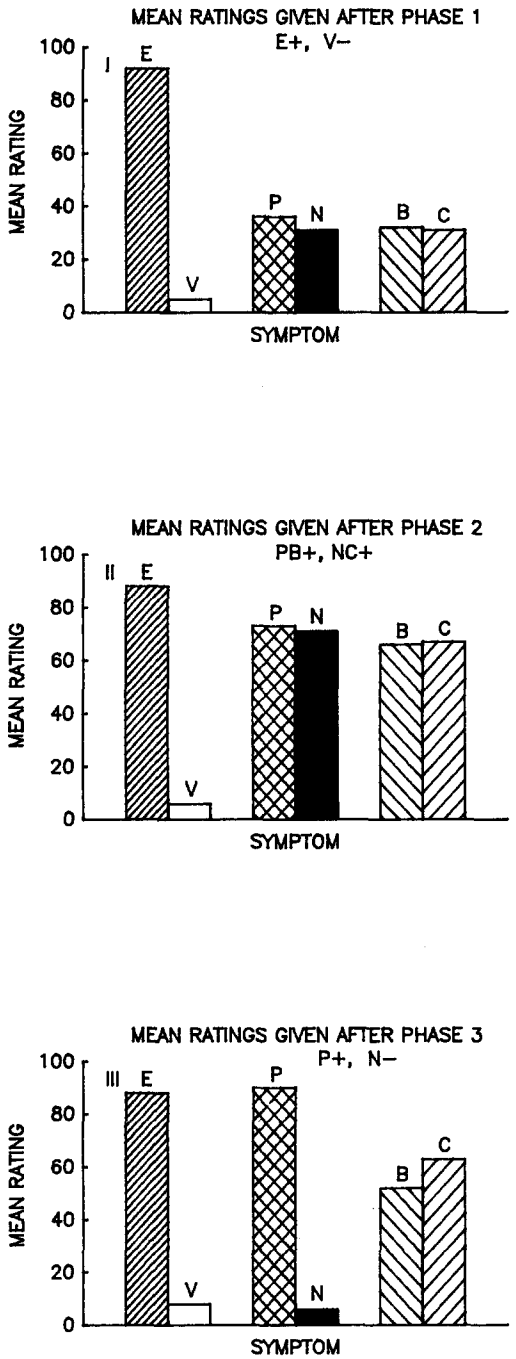

Figure 3. Mean ratings given to each symptom after each of the three phases of Experiment 3. (In Phase 1, symptom E was presented as a good predictor of the disease, whereas symptom V was not predictive. In Phase 2, P was presented in compound with B, and N was presented in compound with C; the disease was present on each compound trial. In Phase 3, symptom N was presented as nonpredictive, whereas symptom P was presented as predictive of the disease.)

Results and Discussion

A considerable number of subjects rated B and C differentially after Phase 1. Because symptoms B and C were not presented in the first phase, such differential ratings signify some preexperimental bias that might interfere with detecting the effect of the experimental manipulation. A criterion was therefore established to ensure that any difference between B and C detected at the end of the experiment was the result of the experimental manipulation and not the result of a preexperimental bias. This criterion also served to reduce the variability of the data. Data from subjects who rated C more than 40 points above or below B at the end of the first phase were eliminated from the analysis. Nine subjects failed to rate B and C similarly at the end of Phase 1, leaving 52 subjects in the primary analysis, 13 in each of the four counterbalance conditions. Primary statistical analyses were conducted using data from these 52 subjects; however, analyses using the data from all subjects yielded almost identical results.[3]

The results for Experiment 3 are shown in Figure 3. The first panel shows the mean ratings given after Phase 1. As expected because of the differential treatment of E and V in Phase 1, E was rated higher than V, $T(52) = 9$, $p < .01$ (51, 1, 0). As seen in previous experiments, symptoms P, N, B, and C were all given moderately positive ratings. Symptom N was rated similarly to P at the end of the first phase, $T(52) = 699.5$, $p > .9$ (24, 28, 0), as expected because neither P nor N had been presented yet. Ratings of C and B did not differ, $T(52) = 635$, $p > .6$ (30, 22, 0), as was the intended result of the previously mentioned criterion.

Panel 2 depicts ratings given after the second phase in which the two compounds were presented. Symptom E was again rated higher than V, $T(52) = 35$, $p < .01$ (49, 3, 0).

[3] When the data from all subjects in Experiment 3 were analyzed, similar results were obtained. After Phase 1, symptom E was rated higher than symptom V (means of 93 and 5, respectively); $T(61) = 10$, $p < .01$ (60, 1, 0). Symptoms P and N were rated similarly (means of 36 and 32, respectively); $T(61) = 896$, $p > .7$ (29, 32, 0). Symptoms B and C were also rated similarly (means of 30 and 34, respectively); $T(61) = 747$, $p > .15$ (37, 24, 0). Note that the criterion for selecting data for the primary analysis was designed to eliminate this slight difference between ratings of B and C. After Phase 2, symptom E continued to be rated higher than V (means of 89 and 6, respectively); $T(61) = 38$, $p < .01$ (58, 3, 0). Symptoms P and N were rated similarly (means of 73 and 70, respectively); $T(61) = 892$, $p > .7$ (30, 31, 0) as were symptoms B and C (means of 65 and 67, respectively); $T(61) = 987$, $p > .7$ (35, 26, 0). After the third phase, symptom E was rated higher than V (means of 89 and 7, respectively); $T(61) = 39$, $p < .01$ (56, 5, 0), and symptom P was rated higher than N (means of 91 and 5, respectively); $T(61) = 27$, $p < .01$ (60, 1, 0). Of primary interest, symptom B was rated lower than C (means of 52 and 64, respectively); $T(61) = 614$, $p < .02$ (34, 27, 0). The difference between the ratings of B and C was larger after Phase 3 (mean = 12) than it was after Phase 2 (mean = 2); $T(45) = 271$, $p < .01$ (30, 15, 16).

Symptoms P and N were rated similarly, $T(52) = 640.5$, $p > .6$ (25, 27, 0). B and C also did not differ from one another, $T(52) = 624$, $p > .5$ (30, 22, 0). Because each of the compounds was predictive of the disease, P, N, B, and C were assigned ratings higher than those given after Phase 1.

The data of primary interest are the ratings given after Phase 3 in which P and N were treated differently. Panel 3 illustrates that as a result of this phase P was rated higher than N, $T(52) = 23$, $p < .01$ (51, 1, 0), demonstrating that the differential treatment of P and N in that phase was effective. Symptom E was once again rated higher than V, $T(52) = 33$, $p, < .01$ (47, 5, 0). Of considerable interest, B was rated reliably lower than C at the end of the third phase, $T(52) = 462.5, p < .05$ (28, 24, 0). This difference constitutes backward blocking, because symptoms B and C were not rated differently until after the differential treatments of P and N. An alternative measure of the backward-blocking effect is to compare the difference in ratings of B and C after Phase 3 to the difference in ratings of B and C after Phase 2. This comparison also yields evidence of backward blocking, because this difference is larger after Phase 3 $T(38) = 181.5$, $p < .01$ (25, 13, 4) (see Footnote 3).

The results of Experiment 3 demonstrate the phenomenon of backward blocking. This effect was numerically smaller than the forward-blocking effect seen in previous experiments. Indeed, only 28 of the 52 subjects rated B less than C after Phase 3. Furthermore, to detect the effect, it was necessary to run a large number of subjects ($N = 52$). The effect was nonetheless clearly present.

This backward-blocking effect implies that subjects retrospectively reevaluated symptoms B and C as a result of new information about symptoms P and N. As further evidence of retrospection, the size of the backward-blocking effect was correlated with memory for which symptoms had appeared in compound together (Spearman $r = .30$, $df = 50$, $p < .05$). That is, subjects who remembered which symptoms had appeared together tended to produce a larger difference between their ratings of B and C. A theory of retrospective processing would anticipate such a correlation because later information about symptom P could not affect the perceived strength of B unless subjects remembered that P and B had appeared in compound together. Such a theory would also predict a correlation between effect size and memory for compounds in Experiments 2, 4, and 5, however. For example, in Experiment 2, subjects with good memory for compounds would be capable of retrospection, and thus should be expected to show a smaller difference between their ratings of the forward- and backward-blocked symptom. This correlation did not appear in any of the other experiments.

As discussed earlier, backward-blocking results have also been obtained by Shanks (1985), who interpreted his results to mean that contingency judgment uses retrospective processing. Like Shanks' results, those of Experiment 3 are inconsistent with associative models such as the Rescorla-Wagner model, which predicts that a backward-blocking procedure should be entirely ineffective in reducing the associative strength of B relative to C. According to this model, the strength of a cue cannot be altered if that cue is not presented; thus, B and C should have been rated identically after Phase 3. By contrast, statistical models predict that blocking will occur with a forward- or backward-trial order because, according to these models, contingency judgment should be perfectly retrospective. These models correctly predict the backward-blocking result of Experiment 3. The joint results of Experiments 2 and 3 indicated that backward blocking does appear (Experiment 3), but that it is of a smaller magnitude than forward blocking (Experiment 2).

Experiment 4

Experiments 4 and 5 investigated another cue-interaction phenomenon: conditioned inhibition. As discussed early in this article, a conditioned inhibitor gains negative predictive strength as a result of compound presentations with a positive predictor when no outcome is present. The Rescorla-Wagner model predicts that conditioned inhibition should develop only with the forward trial order. That is, a potential conditioned inhibitor will gain negative strength only if its partner stimulus is established as a positive predictor before compound presentations. Strengthening the partner stimulus only after compound presentations will not result in the accrual of negative strength. By contrast, statistical models predict that conditioned inhibition will develop with either forward- or backward-trial orders.

Experiment 4 was designed to compare forward- and backward-conditioned inhibition. As such, it contained three phases (see Table 4). In the first phase, symptom P was established as a good predictor of the disease. In the second phase, two symptom compounds (PF and NB) were presented in the absence of the disease. In the third phase, N was presented as a positive predictor. Each of the first three phases also contained presentation of another symptom, earache (designated as E), that was a good predictor of the disease. This good predictor was included so that Phase 2 would contain some presentations of the disease.

This experimental design treated F as a forward inhibitor because symptom P was established as a good predictor before compound presentations with F. By comparison, B was treated as a backward inhibitor because N was established as a good predictor only after compound presentations with B. Associative models predict that F will be rated as more negative than B. Statistical models predict that F and B will be rated similarly.

Method

Subjects and apparatus. The subjects were 68 undergraduates from area colleges. They were paid $5 per hour for this experiment

Table 4
Design for Experiment 4

Phase 1	Phase 2	Phase 3
P → disease	PF → no disease	
	NB → no disease	N → disease
E → disease	E → disease	E → disease

Note. On completion of each of the three phases, subjects were asked to rate each of the five symptoms.

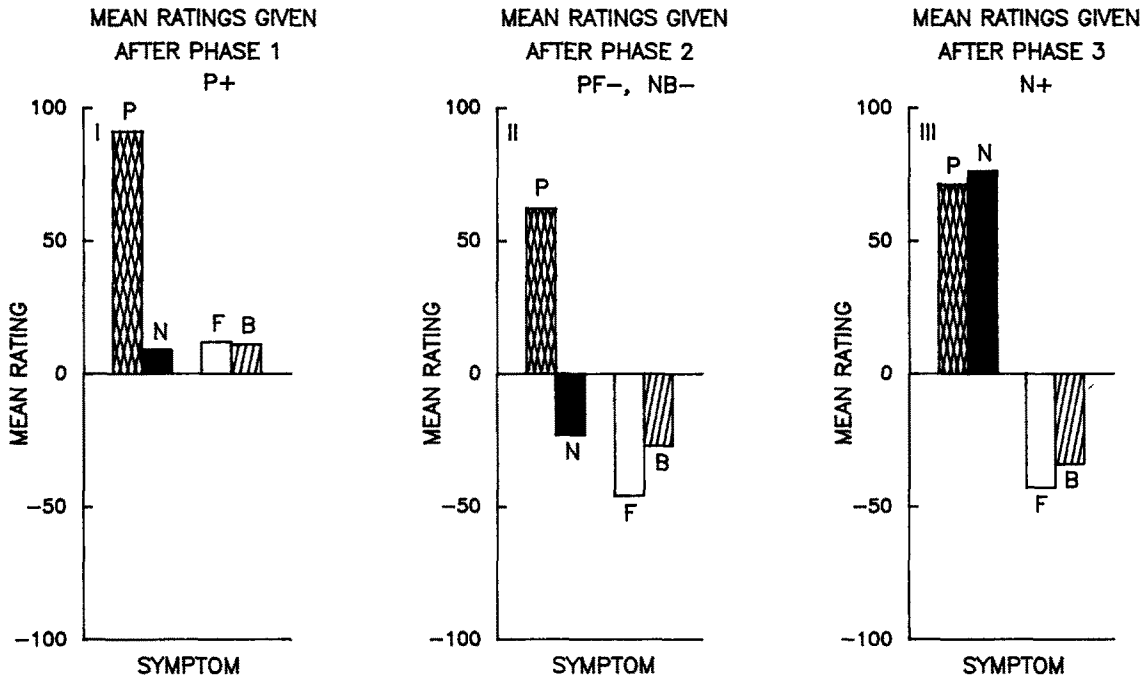

Figure 4. Mean ratings given to each symptom after each of the three phases of Experiment 4. (In Phase 1, symptom P was presented as a good predictor of the disease. In Phase 2, P was presented in compound with F, and N was presented in compound with B; the disease was absent on each compound trial. In Phase 3, symptom N was presented as predictive of the disease.)

and a number of other unrelated experiments. The same computers and program were used as in the previous experiments, with the exception of the particular trial types.

Procedure. The instructions given were similar to those in Experiment 1, as were the details of trial presentation. Each subject was assigned to one of four conditions of a Latin square to counterbalance which symptom names corresponded to the stimulus roles designated as F, B, P, and N. The symptom names were asthma, bruises, coughing, dizziness, and earache.

Because this experiment involved a negatively predictive symptom, the instructions about rating the symptoms were altered. For the end-of-phase ratings, subjects were asked to rate to what extent the presence of each symptom changed the probability that a patient suffered from the disease. Each rating was expressed as a number between −100 and +100, and subjects were instructed to select a positive number if they thought the presence of a particular symptom made it more likely that a patient had morolis. They were instructed to select a negative number if they thought the presence of a symptom made it less likely that the patient had morolis, and to pick a number near zero if they thought that the presence of a symptom did not affect the likelihood that the patient had morolis. Estimates given on each trial, including the test trials, however, used the 0 to 100 scale as in previous experiments.

Experiment 4 consisted of three phases. During the first phase, one symptom, earache, was displayed on 12 trials on which the disease was present (E+). Symptom P was displayed on 12 other trials in which the disease was present (P+). On an additional 12 trials, no symptoms or disease were present. In the second phase, two types of compound trials were presented. On 9 trials, symptoms P and F were presented and the disease was absent (PF−). On 9 other trials, symptoms N and B were displayed and the disease was absent (NB−). There were also 9 trials on which only symptom E was present and the disease was present (E+). An additional 9 trials displayed no symptoms or disease. In the third phase, there were 12 trials in which E was displayed and the disease was present (E+). On an additional 12 trials, symptom N was displayed and the disease was present (N+). There were also 12 trials displaying no symptoms or disease.

Results and Discussion

A criterion like that used in Experiment 3 was employed in this experiment. Data from subjects who rated F more than 40 points above or below B at the end of the first phase were eliminated from the analysis. Eight subjects failed to rate B and F similarly at the end of Phase 1, leaving 60 subjects in the primary analysis, 15 in each of the four counterbalance conditions. An analysis of the data from all 68 subjects yielded similar results.[4]

Figure 4 shows the mean ratings given for symptoms P, N, F, and B. Ratings of symptom E are not displayed but

[4] Results from all the subjects in Experiment 4 showed a similar pattern. After Phase 1, P was rated higher than N (means of 91 and 9, respectively); $T(68) = 18$, $p < .01$ (64, 4, 0). Symptom F was rated similarly to B (means of 15 and 11, respectively); $T(68) = 1270$, $p > .5$ (35, 33, 0). After Phase 2, P continued to be rated higher than N (means of 56 and −23, respectively); $T(68) = 48$, $p < .01$ (62, 6, 0). Symptom F was rated lower than symptom B (means of −50 and −26, respectively); $T(68) = 511$, $p < .01$ (49, 19, 0). After Phase 3, symptom P was rated similarly to symptom N (means of 65 and 74, respectively); $T(68) = 1014$, $p > .3$ (36, 32, 0). Symptom F continued to be rated lower than symptom B (means of −40 and −29, respectively); $T(68) = 710$, $p < .01$ (47, 21, 0).

remained high throughout the experiment. Panel 1 shows that after the first phase symptom P was rated higher than N, $T(60) = 6.5$, $p < .01$ (58, 2, 0), as expected. Symptoms F and B were rated similarly, $T(60) = 946$, $p > .8$ (32, 28, 0), as was the intended result of the criterion mentioned previously. As seen in other experiments, symptoms that had not previously been presented (N, F, and B) were given slightly positive ratings.

The second panel depicts the ratings given after Phase 2. Symptom P continued to be rated higher than N, $T(60) = 35$, $p < .01$ (47, 5, 0). Symptom F was rated lower than B, $T(60) = 436$, $p < .01$ (34, 18, 0), providing evidence of forward inhibition. Symptom F had been presented in compound with a good predictor, whereas B had been presented with a nonpredictive symptom. As a result, F acquired negative predictive strength, as indicated by the fact that it was rated lower than B. It is interesting to note that although symptoms B and N would be expected to have no predictive strength, they were given moderately negative ratings. Other research (Chapman & Robbins, 1990) also showed that stimuli that are presented in the absence of an outcome tend to be rated as slightly negative. This phenomenon was not observed in Experiments 1 to 3 because subjects were not permitted to use negative numbers as ratings.

The third panel shows the data of primary interest. After Phase 3, in which N was presented as a good predictor, N was rated similarly to P, which had previously been established as a good predictor, $T(60) = 838$, $p > .5$ (23, 29, 0). Despite the change in relative ratings of P and N, F was nonetheless rated lower than B, $T(60) = 554$, $p < .01$ (33, 19, 0). This difference indicates that forward inhibition is more effective than backward inhibition.

An additional comparison demonstrated that the difference between F and B was smaller after the third phase than it was after the second phase $T(45) = 340.5$, $p < .05$ (30, 15, 7). This decrease in the difference between F and B suggests that B did acquire some backward inhibition, although such a decrease could also result from a modification in scale use after the third phase. Nonetheless, the forward-inhibition procedure was demonstrably more effective than the backward procedure (see Footnote 4).

Data from the ratings given after Phase 3 indicate that the phenomenon of conditioned inhibition is affected by trial order. Specifically, the forward-trial order was more effective in obtaining inhibition than was the backward-trial order. These results are similar to those seen in Experiment 2, which showed that blocking was affected by trial order. However, the results of Experiment 4 expand the scope of trial-order effects. It appears that multiple types of cue-interaction phenomena are subject to the order in which trials are presented. The trial-order effects obtained in this experiment and in Experiment 2 are consistent with associative models such as the Rescorla-Wagner model. They cannot be explained by statistical models, which posit no effect of trial order.

Experiment 5

As with the blocking phenomenon, associative models not only predict that a forward-conditioned inhibition procedure should produce more cue interaction than a backward-conditioned inhibition procedure, but also that a backward-conditioned inhibition procedure should produce no cue interaction at all. Experiment 5 sought to further establish the similarity between blocking and conditioned inhibition by investigating whether a backward-conditioned inhibition procedure produces any cue interaction.

The design of Experiment 5 (shown in Table 5) provided a comparison between a stimulus embedded in a backward-conditioned inhibition procedure and a control stimulus. Two pairs of symptoms (PB and NC) were presented and were followed by the absence of the disease. In a later phase of the experiment, symptom P was presented as a good predictor of the disease, whereas N was not. Thus, symptom B was embedded in a backward-conditioned inhibition design because its partner (symptom P) was established as a good predictor after the compound phase. Symptom C served as a control stimulus because its partner (symptom N) was not established as a good predictor of the disease.

To make the design of Experiment 5 analogous to that of Experiment 4, an initial phase preceded the compound presentations. This phase involved two other symptoms (earache and fever) that were both presented as good predictors of the disease. In addition, trials in which earache was followed by the disease were included in the second and third phases to ensure that the disease was occasionally presented in the compound phase.

If this backward-conditioned inhibition design yields cue interaction, then symptom B will be assigned a lower (more negative) rating than C. Statistical models predict such a result. If, as associative models predict, a backward-conditioned inhibition procedure is ineffective, symptoms B and C should be rated similarly.

Method

Subjects and apparatus. The subjects were 51 undergraduates from area colleges. They were paid $5 per hour for this experiment and a number of other unrelated experiments. The same computers and program were used as in the previous experiments, with the exception of the particular trial types.

Procedure. The instructions were the same as those in Experiment 4, as were the details of trial presentation and elicitation of ratings. Each subject was assigned to one of four conditions of a Latin square to counterbalance which symptom names corresponded to the stimulus roles designated as B, C, P, and N. The symptom names were those used in Experiment 1.

Experiment 5 consisted of three phases. During the first phase, two symptoms, earache and fever, were each displayed on 12 trials on which the disease was also present (E+ and V+). On an additional

Table 5
Design for Experiment 5

Phase 1	Phase 2	Phase 3
V → disease	PB → no disease NC → no disease	P → disease
E → disease	E → disease	E → disease

Note. On completion of each of the three phases, subjects were asked to rate each of the six symptoms.

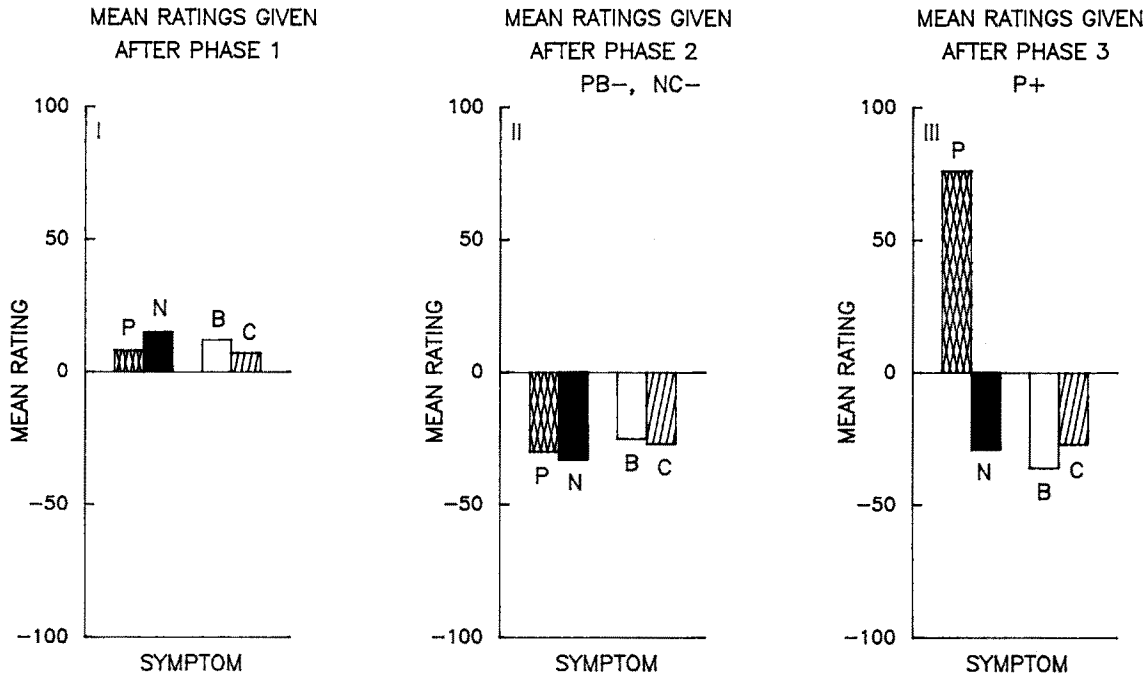

Figure 5. Mean ratings given to each symptom after each of the three phases of Experiment 5. (In Phase 1, symptoms E and V (not displayed) were presented as good predictors of the disease. In Phase 2, P was presented in compound with B, and N was presented in compound with C; the disease was absent on each compound trial. In Phase 3, symptom P was presented as predictive of the disease.)

12 trials, no symptoms or disease were present. In the second phase, two types of compound trials were presented. On 9 trials, symptoms P and B were presented and the disease was absent (PB−). On 9 other trials, symptoms N and C were displayed and the disease was absent (NC−). There were also 9 trials on which only symptom E was present and the disease was present (E+). An additional 9 trials displayed no symptoms or disease. In the third phase, there were 12 trials in which E was displayed and the disease was present (E+). On an additional 12 trials symptom P was displayed and the disease was present (P+). There were also 12 trials displaying no symptoms or disease.

Results and Discussion

A criterion such as that used in Experiment 3 was employed in this experiment. Data from subjects who rated B more than 40 points above or below C at the end of the first phase were eliminated from the analysis. Seven subjects failed to rate B and C similarly at the end of Phase 1, leaving 44 subjects in the primary analysis, 11 in each of the four counterbalance conditions. An analysis of the data from all 51 subjects yielded the same results.[5]

Figure 5 shows the mean ratings given for symptoms P, N, B, and C. Ratings of symptom E and V are not displayed but remained high throughout the experiment. Panel 1 shows ratings given after the first phase. Surprisingly, symptom P was actually rated somewhat lower than N, $T(44) = 313.5$, $p < .05$ (30, 14, 0). B and C were rated similarly, $T(44) = 516$, $p > .80$ (22, 22, 0), as was the intended result of the criterion mentioned previously. As seen in other experiments, symptoms that had not previously been presented (N, C, and B) were given slightly positive ratings.

The second panel depicts the ratings given after the second phase. As seen in Experiment 4, symptoms that had been presented in the absence of the disease received moderately negative ratings. Symptoms P and N were rated similarly, $T(44) = 584$, $p > .30$ (19, 25, 0). Symptoms B and C were also rated similarly, $T(44) = 434$, $p > .40$ (23, 21, 0).

The third panel shows the data of primary interest. After the third phase, in which P was presented as a good predictor, P was rated higher than N, $T(44) = 9.5$, $p < .01$ (41, 3, 0). As a result of the change in relative ratings of P and N, B was rated reliably less than C, $T(44) = 311$, $p < .05$ (27, 17, 0). Thus, these ratings demonstrate evidence of backward-conditioned inhibition, because endowing P with predictive

[5] Results from all the subjects in Experiment 5 showed a similar pattern. After Phase 1, symptom P was rated similarly to symptom N (means of 7 and 9, respectively); $T(51) = 825$, $p > .1$ (12, 33, 0). Symptom B was rated similarly to symptom C (means of 10 and 4, respectively); $T(51) = 801$, $p > .1$ (23, 28, 0). After Phase 2, symptoms P and N continued to be rated similarly (means of −33 and −33, respectively); $T(51) = 715$, $p > .6$ (24, 27, 0). Symptoms B and C were also rated similarly (means of −30 and −30, respectively); $T(51) = 629$, $p > .7$ (26, 25, 0). After Phase 3, symptom P was rated higher than N (means of 76 and −28, respectively); $T(51) = 13.5$, $p < .01$ (47, 4, 0). As a result, symptom B was rated lower than symptom C (means of −34 and −27, respectively); $T(51) = 429$, $p < .05$ (32, 19, 0).

strength after the compound trials induced negative ratings of B (see Footnote 5).

These results indicate that conditioned inhibition is similar to blocking in that a backward-trial order does produce some cue interaction. Like backward blocking, this result is inconsistent with associative models but quite consistent with statistical models.

General Discussion

The studies reported here provide evidence for the importance of trial order in determining judgments of predictive strength. Experiment 1 illustrated the cue-interaction phenomenon known as blocking. Experiment 2 demonstrated the effect of trial order on the blocking phenomenon. Specifically, a forward-blocking procedure was found to be more effective than a backward-blocking procedure in producing cue competition. Experiment 3 illustrated, however, that a backward-blocking procedure does produce some cue competition. Experiment 4 produced a result similar to that obtained in Experiment 2; trial order was found to influence conditioned inhibition, a related cue-interaction phenomenon. The forward-trial order produced more inhibition than did the backward order. In Experiment 5, a backward-conditioned inhibition procedure resulted in some cue interaction.

These results of Experiments 2 and 4 provide support for a class of models that predict an effect of trial order. Associative models such as the Rescorla-Wagner model postulate that the predictive strength of a cue is updated on each trial in which it is presented; thus, trial order can affect the terminal strengths. Statistical models, such as a multiple linear regression, do not predict trial-order effects; thus, they cannot explain the results of Experiments 2 and 4. By contrast, Experiments 3 and 5 point to a failure of associative models to deal with small retrospective effects, which can be explained by statistical models.

As discussed at the beginning of the discussion of Experiment 3, other contingency judgment research has demonstrated similar retrospective results. Shanks (1985) presented subjects with a video game in which they fired shells at passing tanks. The tanks, which traversed a mine field, could explode either because they were hit by shells or because of a mine. Shanks presented subjects with both forward- and backward-blocking procedures. In the forward-blocking procedure, subjects observed the frequency with which the tank exploded as a result of the mines before they had the opportunity to fire shells at the tanks. In the backward-blocking procedure, the observation period followed the firing period. In control procedures, subjects did nothing during the time they would otherwise be observing the effectiveness of the mines. Thus, each subject was successively presented with two blocking sessions and two control sessions. After each session, subjects rated the effectiveness of the shells. Shanks (1985) found that both forward- and backward-blocking procedures produced lower ratings of the shell's effectiveness than did the respective control procedures. Furthermore, the backward-blocking procedure appeared just as effective as the forward-blocking procedure in producing this effect. Thus, Shanks' backward-blocking result is similar to the result obtained in Experiment 3. Unlike the results of Experiment 2, however, Shanks's results indicated that forward- and backward-blocking procedures have comparable effects on predictive strength.

The discrepancy between Shanks's data and the results of Experiment 2 might be explained by procedural differences. First, subjects in Experiment 2 were asked to compare directly a forward-blocked stimulus to a backward-blocked stimulus; the two stimuli appeared in the same phase of the experiment, and subjects assigned ratings to them at the same time. By contrast, Shanks's subjects experienced a backward-blocked stimulus in a different condition than the forward-blocked stimulus. Subjects rated both forward- and backward-blocked shells, but they were not directly asked to compare the two. A request for a direct comparison may have encouraged subjects to reveal the judged difference in predictive strengths.

A second procedural difference that may explain the discrepancy between Shanks's (1985) data and the results of Experiment 2 involves the type of relationship that subjects were asked to judge. In Shanks's experiments, subjects judged the relation between the response of firing a shell and the outcome of the tank explosion. By contrast, subjects in the experiments reported here judged the relation between a symptom and a disease. These judged relationships differ in two respects. First, Shanks's subjects judged the relationship between an act and an outcome, whereas subjects in the experiments reported here judged the relationship between a stimulus and an outcome. Act–outcome judgments may use a different mechanism than stimulus–outcome judgments. Second, in Shanks's experiments, subjects likely viewed the act of firing a shell as a potential cause of the tank explosion. Appropriately enough, Shanks and his colleagues (Dickinson & Shanks, 1985; Shanks, 1989) considered their findings relevant to the study of causality judgments. The judged relationships in the studies reported here do not involve the same sort of causal link. A symptom does not cause a disease; on the contrary, it is reasonable to think that the disease produces the symptom. Therefore, the initial event does not cause the outcome but is merely predictive of it. It is possible that the nature of the causal relation between predictor and outcome may influence the mechanism underlying contingency judgment (see Waldmann & Holyoak, 1990, for a discussion of this issue).

For example, causality judgments may have more of a tendency to use retrospective processing than do predictability judgments. People may be inclined to view only one of the two copresent events as the cause of a later outcome (Kelley, 1973). In the backward-blocking condition, subjects initially had no basis for deciding whether the shells or the mines were the cause of the explosions; however, they later received information indicating that the mines caused explosions. Given that there can be only one cause, concluding that the mines were a cause necessitates the inference that the shells were not the cause. This tendency would result in a large backward-blocking effect, perhaps equal in magnitude to the forward-blocking effect. By contrast, people may be much more inclined to admit multiple predictors of an outcome. Thus, in the backward-blocking procedure, later information about one symptom does not necessitate much reevaluation of other symptoms. Thus, a backward-blocking effect would

be slight, if it occurred at all, and smaller than the forward-blocking effect.

Interestingly, results indicating retrospective processing have also been obtained with animals. In an experiment by Matzel, Schachtman, and Miller (1985), rat subjects were presented with a tone and light in compound followed by a shock. Later presentation of the tone followed by no shock resulted in an increase in fear to the light. Similar results have been obtained by Kaufman and Bolles (1981) and could be easily explained by a statistical model. However, a large body of animal-conditioning literature reports results quite the opposite of those found by Matzel et al. (1985) and Kaufman and Bolles (1981). A number of experiments have demonstrated that, after compound presentations of two cues, altering the strength of one element of the compound results in a similar change in strength of the other element of the compound (Rescorla, 1982; Rescorla & Colwill, 1983; Rescorla & Durlach, 1981; Rescorla & Freberg, 1978). In other words, later treatment of one element of a compound does alter the predictive strength of the other element, but in a direction opposite of that predicted by a statistical model. These findings illustrate that results indicative of retrospective processing in animals are rare.

Although the retrospective effects found here appear to be quite small in magnitude, they are not without precedent. The fact that they have been demonstrated several times suggests that it is worthwhile to consider how they might be incorporated into a model of contingency judgment.

Neither an associative nor a statistical model can deal with the full pattern of results reported here. Associative models, as discussed so far, predict trial-order effects but not retrospective effects. Conversely, statistical models predict retrospective effects but not trial-order effects. A successful model must predict a small amount of retrospection while maintaining a prediction of trial order. A particular version of an associative model may be able to explain this pattern of data. The class of associative models includes a wide variety of adaptive network models. Particular features of some of these models may provide an account of this pattern of results.

As was noted earlier, a network model, like other associative models, examines information one trial at a time, leading to a prediction of trial-order effects. Interestingly, the importance of this prediction about trial-order effects has only recently been recognized by network theorists. McCloskey and Cohen (1989) noted the importance of trial order when applying network models to the problem of sequential learning, learning which involves the successive presentation of multiple sets of training items such that the training sets are processed one at a time. This network application simulates learning tasks in which subjects do not experience pieces of information in a perfectly intermixed manner. They found that a network model given sequential training yielded a solution that was influenced by the trial order used. Specifically, later learning catastrophically interfered with earlier learning. In contrast, if trials from all training sets were intermixed, learning about all the items proceeded without interference. McCloskey and Cohen (1989) viewed this influence of trial order as a flaw or limitation of network models. They regarded the interference, which networks predict will result from sequential learning, as much more extensive than the interference observed with human subjects in learning experiments. However, the results reported here encourage a different view. The ability of a network to predict the effects of trial order might be considered an asset, because it enables the network to explain trial-order effects such as those reported here.

In another series of network simulations, Ratcliff (1990) compared sequential learning to a condition involving intermixed trials. He argued that because many learning tasks provide little opportunity for additional exposure to trials presented early in training, it is important to examine the network solution to a sequential learning task. Like the results of McCloskey and Cohen (1989), Ratcliff's results indicated that in sequential learning simulations the learning of later items drastically interfered with performance to items learned earlier. Such interference was not produced with intermixed training; thus, the order of trial presentation affected the final network weights.

The particular order in which trials are processed could be viewed as controlled exclusively by the learning environment. Alternatively, the order in which trials are processed may be partially controlled by the subject. Ratcliff (1990) designed networks that modeled a human subject rehearsing each trial or group of trials after it was presented. These network models not only processed the trials that were actually presented to the subject; they also rehearsed or recycled previous trials. For example, if a subject was presented with three trials (1, 2, 3), to simulate rehearsal the network model would be presented with each trial multiple times (e.g., 1, 1, 1, 2, 2, 2, 3, 3, 3) or with each group of trials multiple times (e.g., 1, 2, 3, 1, 2, 3, 1, 2, 3). The particular method of simulating rehearsal affects the network solution, because it influences the order in which the network processes trials.

Ratcliff's (1990) notion of rehearsal could be used to explain the results reported here. If subjects occasionally rehearsed trials from previous phases, the retrospective effects found in Experiments 3 and 5 could be explained. This rehearsal could be accomplished in a variety of ways. For instance, as Ratcliff (1990) suggested, after each trial the network might recycle a small constant number of previous trials. In Experiment 3, after the first few trials of Phase 3, the subject may have rehearsed some of the trials from Phase 2. After symptom P gained a small amount of strength during the first few trials of Phase 3, rehearsal of one of the PB+ trials would result in a reduction in symptom B's predictive strength.

As an alternative conceptualization of rehearsal, presentation of a trial containing a particular symptom might retrieve memories of other trial types that contained that same symptom. For example, in Phase 3 of Experiment 3, presentation of a P+ trial could have retrieved memories of other trials containing symptom P, that is, PB+ trials. If subjects rehearsed a few Phase 2 trials during Phase 3, the strength of the backward-blocked symptom (B) would become slightly less than that of the control symptom (C). A small amount of rehearsal would be sufficient to produce a small backward-blocking effect. A similar account could explain the backward-conditioned inhibition obtained in Experiment 5.

Ratcliff (1990) suggested that such storage and rehearsal of remote previous trial types would need to be accomplished

by some mechanism separate from the network itself (perhaps by a second network), because a network stores the predictive strengths of each stimulus but does not store representations of trial types. However, a particular type of network model, such as one proposed by McClelland and Rumelhart (1985), could be capable of both storing predictive strengths and accomplishing a process resembling rehearsal. In McClelland and Rumelhart's model, all the nodes within a layer are connected to one another as well as to nodes from the layers. Because of the interconnections among the input nodes, if a partial pattern is presented to the input layer, the network can complete the rest of the pattern. This process is similar to the idea that the presentation of a trial containing one symptom retrieves a memory of other trials containing that symptom. For example, a P+ trial in the third phase of Experiment 3 might act as a partial pattern, which activates the full PB+ pattern.

With extensive amounts of rehearsal, the network solution will approach the result predicted by statistical models. Limited rehearsal, however, would still allow effects of trial order. Thus, subjects may have engaged in rehearsal in Experiments 2 and 4 as well. Such rehearsal would have reduced the magnitude of the order effects observed in those experiments; however, with only a small amount of rehearsal, the order effect would still be expected to occur.

A particular type of associative model is therefore capable of explaining the pattern of results reported here. Associative models, such as network models and conditioning models, process information in a trial-by-trial manner; thus, they predict effects of trial order. These models can simulate incomplete retrospective processing by limited rehearsal of at least some trials.

An interesting comparison can be made between the work discussed here and experiments reported by Koh and Meyer (1991) in which subjects learned a relationship between the length of a line and the temporal duration of a response. The results were best explained by an adaptive regression model, which postulates that subjects perform a combination of log-linear and polynomial regression analysis. Koh and Meyer's experiments appear to contradict the results presented here, which demonstrate that statistical models, such as regression, do not provide a good account of judgments about predictive relationships. It is important to note, however, that these two sets of results are not as contradictory as they appear. Statistical models and associative models differ only in their predictions about the effect of trial order. If many intermixed trials are presented, as in the Koh and Meyer experiments, the two classes of models yield identical predictions (Gluck & Bower, 1988); therefore, Koh and Meyer's results do not distinguish between the statistical and associative models presented here.

An interesting aspect of the Koh and Meyer (1991) results is that the log-linear and polynomial regression provided a better account of the results than did linear regression. Associative models such as the Rescorla-Wagner model or the Gluck and Bower (1988) network model are most similar to linear regression. Because most of the experiments supporting these models, including those presented here, used only binomial cues (i.e., a symptom could be present or absent), they cannot distinguish between linear and nonlinear models. The Koh and Meyer results suggest that the most successful associative model may include a nonlinear component. Such nonlinearity could be accomplished, for example, with the use of intermediate layers of hidden nodes in a network model.

To summarize, statistical methods are generally accepted as the normative approach to contingency judgment. As evidenced by the results reported here, these normative statistical models do not provide a good descriptive model of judgments about predictive relationships. Associative models provide a more accurate description of how people judge contingencies. Associative models can be viewed as approximations of statistical tests; however, they differ from statistical techniques in that, according to associative models, predictive strengths are affected by the order in which information is received. These order effects represent an important deviation from the normative statistical models.

The data reported here indicate how sequential information is processed. These results suggest that subjects update their ratings of predictive stimuli on a trial-by-trial basis while also occasionally rehearsing some past trials. This cognitive process can be simulated by an associative model that rehearses trials a limited number of times.

References

Alloy, L. B., & Abramson, L. Y. (1979). Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of Experimental Psychology: General, 108,* 441–485.

Beyth-Marom, R. (1982). Perception of correlation reexamined. *Memory & Cognition, 10,* 511–519.

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

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

Crocker, J. (1981). Judgment of covariation by social perceivers. *Psychological Bulletin, 90,* 272–292.

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.

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

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

Holyoak, K. J., Koh, K., & Nisbett, R. E. (1990). A theory of conditioning: Inductive learning within rule-based default hierarchies. *Psychological Review, 96,* 315–340.

Kamin, L. J. (1968). 'Attention-like' processes in classical conditioning. In M. R. Jones (Ed.), *Miami symposium on the production of behavior: Aversive stimulation* (pp. 9–33). Coral Gables, FL: University of Miami Press.

Kamin, L. J. (1969). Predictability, surprise, attention, and condi-

tioning. In B. Campbell & R. Church (Eds.), *Punishment and aversive behavior* (pp. 279–296). New York: Appleton.

Kaufman, M. A., & Bolles, R. C. (1981). A nonassociative aspect of overshadowing. *Bulletin of the Psychonomic Society, 18,* 318–320.

Kelley, H. H. (1973). The process of causal attribution. *American Psychologist, 28,* 107–128.

Koh, K., & Meyer, D. E. (1991). Function learning: Induction of continuous stimulus-response relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 811–836.

Mackintosh, N. J. (1975a). Blocking of conditioned suppression: Role of the first compound trial. *Journal of Experimental Psychology: Animal Behavior Processes, 1,* 335–345.

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

Matzel, L. D., Schachtman, T. R., & Miller, R. R. (1985). Recovery of an overshadowed association achieved by extinction of the overshadowing stimulus. *Learning and Motivation, 16,* 398–412.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114,* 159–188.

McClelland, J. L., Rumelhart, D. E., & the PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 2: Psychological and biological models.* Cambridge, MA: MIT Press.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory, 24* (pp. 109–165). New York: Academic Press.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General, 117,* 68–85.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review, 90,* 339–363.

Pavlov, I. P. (1927). *Conditioned reflexes.* Oxford, England: Oxford University Press.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review, 67,* 532–552.

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68,* 29–46.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review, 97,* 285–308.

Rescorla, R. A. (1969). Pavlovian conditioned inhibition. *Psychological Bulletin, 72,* 77–94.

Rescorla, R. A. (1981). Within-signal learning in autoshaping. *Animal Learning and Behavior, 9,* 245–252.

Rescorla, R. A. (1982). Some consequences of associations between the excitor and the inhibitor in a conditioned inhibition paradigm. *Journal of Experimental Psychology: Animal Behavior Processes, 8,* 288–298.

Rescorla, R. A., & Colwill, R. M. (1983). Within-compound associations in unblocking. *Journal of Experimental Psychology: Animal Behavior Processes, 9,* 390–400.

Rescorla, R. A., & Durlach, P. J. (1981). Within-event learning in Pavlovian conditioning. In N. E. Spear & R. R. Miller (Eds.), *Information processing in animals: Memory mechanisms* (pp. 81–111). Hillsdale, NJ: Erlbaum.

Rescorla, R. A., & Freberg, L. (1978). The extinction of within-compound flavor associations. *Learning and Motivation, 9,* 411–427.

Rescorla, R. A., & Holland, P. C. (1977). Associations in Pavlovian conditioned inhibition. *Learning and Motivation, 8,* 429–447.

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 theory and research* (pp. 64–99). New York: Appleton-Century-Crofts.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations.* Cambridge, MA: MIT Press.

Shaklee, H., & Tucker, D. (1980). A rule analysis of judgments and covariation between events. *Memory & Cognition, 8,* 459–467.

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

Shanks, D. (1986). Selective attribution and the judgment of causality. *Learning and Motivation, 17,* 311–334.

Shanks, D. (1989). Selectional processes in causality judgments. *Memory & Cognition, 17,* 27–34.

Shanks, D., & Dickinson, A. (1987). Associative accounts of causality judgment. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory, 21* (pp. 229–261). New York: Academic Press.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill.

Stone, G. O. (1986). An analysis of the delta rule and the learning of statistical association. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations* (pp. 444–459). Cambridge MA: MIT Press.

Waldmann, M. R., & Holyoak, K. J. (1990). Can causal induction be reduced to associative learning? (Tech. Rep. No. UCLA-CRSP-90–6). Los Angeles: University of California, Los Angeles.

Ward, W. C., & 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.

Wasserman, E. A., Chatlosh, D. L., & Neunaber, D. J. (1983). Perceptions of causal relations in humans: Factors affecting judgments of response-outcome contingencies under free-operant procedures. *Learning and Motivation, 14,* 406–432.

Received August 6, 1990
Revision received December 12, 1990
Accepted January 26, 1991 ■