Journal of Experimental Psychology: General

Inferential Dependencies in Causal Inference: A Comparison of Belief-Distribution and Associative Approaches

Christopher D. Carroll, Patricia W. Cheng, and Hongjing Lu

Online First Publication, September 10, 2012. doi: 10.1037/a0029727

Journal of Experimental Psychology: General

0096-3445/12/$12.00 DOI: 10.1037/a0029727

Inferential Dependencies in Causal Inference: A Comparison of Belief-Distribution and Associative Approaches

Christopher D. Carroll, Patricia W. Cheng, and Hongjing Lu
University of California, Los Angeles

Causal evidence is often ambiguous, and ambiguous evidence often gives rise to *inferential dependencies*, where learning whether one cue causes an effect leads the reasoner to make inferences about whether other cues cause the effect. There are 2 main approaches to explaining inferential dependencies. One approach, adopted by Bayesian and propositional models, distributes belief across multiple explanations, thereby representing ambiguity explicitly. The other approach, adopted by many associative models, posits *within-compound associations*—associations that form between potential causes—that, together with associations between causes and effects, support inferences about related cues. Although these fundamentally different approaches explain many of the same results in the causal literature, they can be distinguished, theoretically and experimentally. We present an analysis of the differences between these approaches and, through a series of experiments, demonstrate that models that distribute belief across multiple explanations provide a better characterization of human causal reasoning than models that adopt the associative approach.

Keywords: causal inference, inferential dependencies, retrospective revaluation, cue competition

Causal evidence is often ambiguous. When trying to identify the cause of a recent illness, the reason why a friend failed to return a phone call, or the cause of a car accident, possible explanations abound. In such situations, subsequent learning about one of the possible causes may support inferences about the other possible causes. Consider, for example, a traveler who becomes ill after a flight where he ate a suspect meal and sat next to a coughing passenger. His illness may have been caused by the meal or by his coughing neighbor. After learning that other passengers who ate the inflight meal did not become ill, the traveler would probably conclude that the cause was his coughing neighbor. In such circumstances, it seems as if the traveler retrospectively reevaluates the ambiguous initial evidence (the two plausible causes of his illness) in light of the subsequent evidence (the harmlessness of the inflight meal). Consequently, the inference is said to involve retrospective revaluation (e.g., Van Hamme & Wasserman, 1994). More generally, we say that there is an *inferential dependency* between two possible causes when learning about one of them can support an inference regarding the other.

How should we explain inferential dependencies in causal reasoning? There are two main approaches to the problem. The associative approach explains inferential dependencies by utilizing *within-compound associations*—associations that form *between* potential causes, in addition to the typical associations between each cause and its effect (e.g., Dickinson & Burke, 1996; Stout & Miller, 2007; Van Hamme & Wasserman, 1994). Within-compound associations are assumed to form when potential causes co-occur, as is typically the case when there is confounding and thus the evidence is ambiguous. The association between co-occurring cues—say, potential causes c_1 and c_2—allows the weight of the association between c_1 and the effect e to be updated for events (trials) on which c_1 is absent; when c_2 occurs, its activation can activate c_1 via the within-compound association. This is unlike in typical associative models, in which only cues that are present are activated and eligible for updating. The within-compound association thereby provides a representation for explaining inferential dependencies in situations involving ambiguity. For example, an associative model might posit that the traveler's meal and the coughing neighbor are associated through a within-compound association. The within-compound association could be used to support the inference that, if one of the cues is not causal, then the other should be. Note, however, that at any given moment, an associative network, regardless of whether it supports within-compound associations, is in a single state where each associative strength is estimated by a single value.

Thus, various alternative explanations of ambiguous evidence would have to map onto the same state of an associative network. In other words, the approach does not provide a means for representing multiple explanations at the same time. For example, consider a series of trials in which two cues, A and B, simultaneously occur and a target effect follows. In an associative network, it seems reasonable to represent this ambiguous evidence by a state where each of the two cues has a cue–effect association with

Christopher D. Carroll and Patricia W. Cheng, Department of Psychology, University of California, Los Angeles; Hongjing Lu, Departments of Psychology and Statistics, University of California, Los Angeles.

The preparation of this article was supported by Air Force Office of Scientific Research Grant FA 9550-08-1-0489 to Alan Yuille and Patricia W. Cheng. Preliminary reports of this research were presented at the 32nd and 33rd Annual Conferences of the Cognitive Science Society. We thank David Danks for extremely helpful comments, and we thank Betty Huang and Aaron Placensia for assistance with data collection.

Correspondence concerning this article should be addressed to Christopher D. Carroll, who is now at Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213. E-mail: cdcarroll@gmail.com

moderate strength. However, consider two explanations of the ambiguous pattern of events: (a) Both cues cause the effect, each with a moderate causal strength, or (b) only one cue causes the effect and that cue causes the effect every time, but it is unknown which cue is the cause. These explanations are distinct, and they imply different inferential dependencies. They also involve different types of uncertainty: The first involves uncertainty regarding the causal mechanism (e.g., information on some enabling conditions is missing, so that the cause leads to the effect only sometimes), and the second involves uncertainty regarding which cue is causal (i.e., causal structure). Yet the representation of the ambiguous evidence by a single network state means that there is only a single cue–effect association for each cue.

Suppose a new trial indicates that introducing Cue A brings about the effect. For the first explanation, this new information from the single trial has no significance; both cues remain candidate causes. But for the second explanation, the same information should lead to a revision of belief: Cue B becomes eliminated as a candidate cause. A single network state cannot capture these different implications.

In contrast, the belief-distribution approach represents multiple explanations of ambiguous evidence and distributes belief across the possible explanations in accordance with the plausibility of each explanation (e.g., Kruschke, 2008). Returning to our traveler example, given that the traveler became ill, a belief-distribution approach may distribute belief across three alternative explanations where the illness is attributed to (a) the inflight meal alone, (b) the coughing neighbor alone, or (c) both the inflight meal and the coughing neighbor.[1] The belief-distribution approach predicts that the reasoner should entertain all three explanations at the same time, each as a distinct possibility with its respective degree of plausibility, in contrast to the conflation of all possible explanations corresponding to the single state of an associative network at one moment.

Belief distribution can be formalized in terms of propositional logic (e.g., De Houwer, Beckers, & Vandorpe, 2005; Lovibond, 2003) or Bayesian inference, where the distribution of belief is captured by probability distributions over alternative hypotheses (Jaynes, 2003; see also Duda, Hart, & Stork, 2000). Because Bayesian models make fine-grained probabilistic predictions that can be quantitatively compared to the predictions of the associative models, we focus on Bayesian rather than propositional models when evaluating the belief-distribution approach. If Bayesian models account for human judgments better than associative models, it would indicate that intuitive inferential dependencies involve belief distribution.

Bayesian models, which exemplify the belief-distribution approach, have been widely used to account for human causal inference (e.g., Griffiths & Tenenbaum, 2005, 2009; Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008), with greater success than the most well-known associative learning model, that of Rescorla and Wagner (R-W) model (Rescorla & Wagner, 1972; for a review, see Holyoak & Cheng, 2011). However, Bayesian models have not been directly compared across multiple paradigms to the more advanced associative models that can explain retrospective revaluation. The comparisons have instead focused on a single associative model (the R-W model) and a few experimental paradigms (mostly forward and backward blocking; e.g., Daw, Courville, & Dayan, 2008; Kruschke, 2008; Lu, Rojas, Becker, & Yuille, 2008; Sobel, Tenenbaum, & Gopnik, 2004). Moreover, because the typical Bayesian model differs from the typical associative model not only in how it represents ambiguous evidence but also along various other dimensions, a direct comparison between two models fails to establish *why* one model might succeed where the other model fails.

Therefore, although we compare specific models that represent the belief-distribution and associative approaches, our goal is not to compare the models per se. Instead, the present article aims to study a fundamental difference between the associative and belief-distribution approaches: namely, the difference between the representation of ambiguous conclusions in each approach. Previous discussions of associative models have not been framed in terms of ambiguous evidence and inferential dependencies, being typically framed in terms of retrospective revaluation instead. We reframe the discussion to reveal a basic and general weakness of the associative approach, with and without within-compound associations. Toward this end, we compare these approaches from a computational perspective and develop empirical tests that clearly differentiate between them. In particular, we consider the implications of these approaches in situations where (a) there is ambiguous evidence, which often creates inferential dependencies, and (b) the evidence is unambiguous but within-compound associations predict inferential dependencies.

Previous research indicates that model predictions may depend on a variation that applies to both Bayesian and associative models—namely, whether the generating function is linear or noisy-logical (i.e., noisy-or and noisy-and-not; Cheng, 1997; Griffiths & Tenenbaum, 2005; Lu, Yuille, et al., 2008). These variants, which reflect different definitions of independence, specify different ways of relating causal structures to observations. We consider both variants of both kinds of models. Thus, our auxiliary goal is to assess the role of the generating function in explaining inferential dependencies.

Computational Approaches to Explaining Inferential Dependencies

We evaluate the belief-distribution and associative approaches by considering a Bayesian model of causal inference, adapted from the proposal of Griffiths and Tenenbaum (2005), and contrasting it with four associative models: the Rescorla-Wagner model (Rescorla & Wagner, 1972) and three advanced associative models. There are significant differences among the three advanced associative models; however, all infer a within-compound association when cues are presented simultaneously and then use the learned within-compound association to establish an inferential dependency between the cues. Additionally, all of the associative models share the assumption that the learner forms and updates a single hypothesis that best "fits" the observed data.

To illustrate the workings of the models, we consider the inferential dependencies created by the ambiguous evidence in variations of *recovery from overshadowing* (e.g., Kaufman & Bolles, 1981; Matzel, Schactman, & Miller, 1985) and *backward blocking* (e.g., Shanks, 1985). In both paradigms, the initial ambiguous evidence shows that the effect occurred following the presentation

[1] The second explanation from two paragraphs earlier corresponds to the possibility that either Explanation a or Explanation b here is correct.

of two cues (AB+). Here and in the rest of this article, the possible causes and effects that we consider are all binary variables with a "present" and an "absent" value. Thus, we denote the potential causes that are present on a trial-type by letters and the presence and absence of the effect by + and –, respectively. In recovery from overshadowing (AB+ A–), the subsequent A– observation shows that one of the two cues (labeled A here) does not cause the effect. In backward blocking (AB+ A+), the subsequent A+ observation shows that one of the cues causes the effect. In studies of Pavlovian conditioning with rats (e.g., Kaufman & Bolles, 1981; Miller & Matute, 1996) and in studies of human causal learning (e.g., Larkin, Aitken, & Dickinson, 1998; Shanks, 1985) with both recovery from overshadowing and backward blocking paradigms, researchers have found that the ambiguous AB+ trials can create inferential dependencies between the cues.

Most of the models predict that when compared with a control condition (AB+), people have a stronger expectation that Cue B causes the effect in recovery from overshadowing and a weaker expectation that it does so in backward blocking. However, the models make different predictions about the degree of uncertainty regarding whether Cue B causes, or does not cause, the effect in recovery from overshadowing and in backward blocking.

A Bayesian Model

Figure 1 illustrates how our Bayesian model explains recovery from overshadowing and backward blocking. When presented with ambiguous AB+ evidence, the model distributes belief across four explanations (see the top row of Figure 1). This distribution of belief implies an inferential dependency between the cues, as revealed by two unequal conditional probabilities: the probability that Cue B causes the effect given that Cue A does *not* cause the effect, $P[B \rightarrow E \mid \sim (A \rightarrow E)] = .25/(.25 + .00) = 1.0$ (see the first and third graphs in the top row of Figure 1), and the probability that Cue B causes the effect given that Cue A *causes* the effect, $P(B \rightarrow E \mid A \rightarrow E) = .50/(.50 + .25) \approx .67$ (see the second and fourth graphs in the top row of Figure 1). These differing probabilities imply that learning whether Cue A causes the effect provides information about whether Cue B causes the effect. The final distribution of belief in recovery from overshadowing (see lower left graphs in Figure 1) and backward blocking (see lower right graphs in Figure 1) illustrates the inferences supported by the inferential dependency.

The Bayesian model that we develop, an extension of Griffiths and Tenenbaum's (2005) model, allows the causal links to be generative or preventive. Like theirs, our model represents each explanation as a *causal graph*. In experiments involving preventive causes, we consider causal graphs where each causal link can be generative, preventive, or nonexistent. In experiments without preventive causes, we only consider the causal graphs where the causal links are either generative or nonexistent. Because we only consider experiments where there are multiple cues and a single effect, we represent a causal graph as a vector of causal links, **l**, letting $l_i = 1$ denote a generative causal relationship between cue i and the effect, $l_i = 0$ denote the absence of a causal relationship, and $l_i = -1$ denote a preventive causal relationship. For n cues, there are 3^n causal graphs to consider when we allow for preventive causation and 2^n causal graphs when we do not. We associate each causal link with a weight that represents the strength of the causal relationship, and we represent these weights as a vector **w**, where $0 \leq w_i \leq 1$ for each w_i. For simplicity, we omit consider-

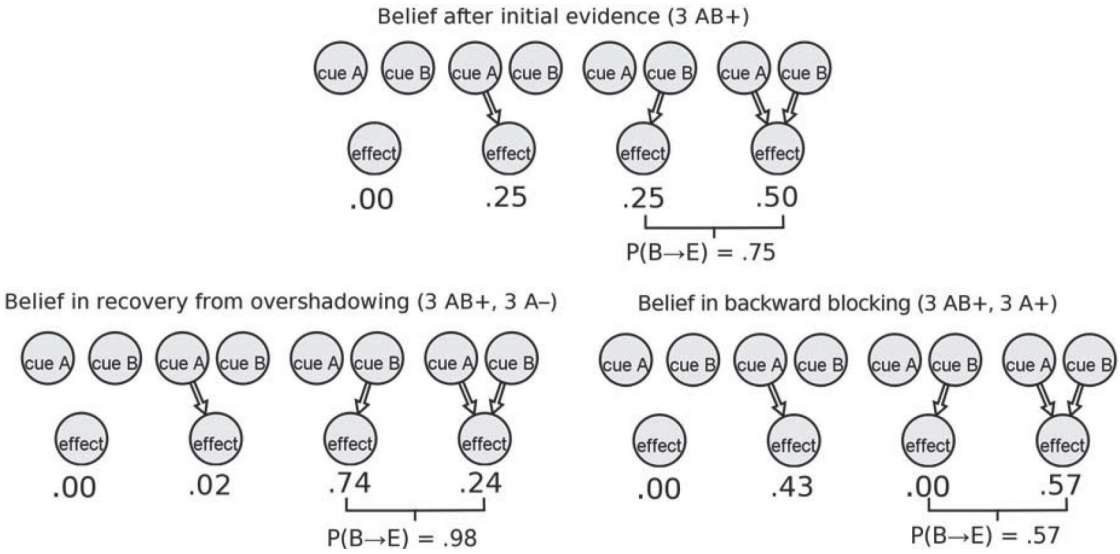

Figure 1. The predictions of the Bayesian model for recovery from overshadowing and backward blocking. Explanations of the data (e.g., 3 AB+) are represented as causal graphs, where arrows represent causal relationships. The causal graph where both cues cause the effect represents an explanation where Cues A and B independently (and not conjunctively) cause the effect. The number below a causal graph represents the posterior probability of the explanation given the data. Observe that the model distributes belief across multiple explanations and that evidence indicating whether Cue A causes the effect influences the model's predictions about whether Cue B causes the effect.

ation of conjunctive causes under the assumption that people only posit conjunctive causes when the observations cannot be explained by simple causes alone (Novick & Cheng, 2004). None of our experiments contain observations of this sort.

To represent a trial, we let the vector **c** denote the presence ($c_i = 1$) or absence ($c_i = 0$) of the cues and let e denote the presence ($e = 1$) or absence ($e = 0$) of the effect. In order to specify the probability of the effect as a function of its causes, we adopt the noisy-or and noisy-and-not generating functions for generative and preventive causation, respectively. These functions are derived from the assumptions of causal power (Cheng, 1997). Let G be the set of indexes such that $l_i = 1$ (i.e., generative causes of e), and let P be the set of indexes such that $l_i = -1$ (preventers of e). Using the noisy-or and noisy-and-not functions, the probability of the effect is

$$P(e = 1|\mathbf{c}, \mathbf{l}, \mathbf{w}) = \left[1 - \prod_{g \in G} (1 - w_g)^{c_g}\right] \prod_{p \in P} (1 - w_p)^{c_p}. \quad (1)$$

Given data D that provides a frequency count $N(e, \mathbf{c})$ for each combination of the presence/absence of the effect and the cues and the assumption that the trials are independent, the likelihood of the data can be written as a function of the causal graph and its weights:

$$P(D|\ \mathbf{w}, \mathbf{l}) = \prod_{(e,\mathbf{c})} P(e|\mathbf{c}, \mathbf{l}, \mathbf{w})^{N(e,\mathbf{c})}. \quad (2)$$

As shown in Equations 3–5, where n is the number of cues and t is the number of possible values for a causal link ($t = 3$ when preventive causes are considered; $t = 2$ otherwise), we assume uninformative prior distributions on **w** and **l:**

$$\begin{aligned} P(\mathbf{w}, \mathbf{l}) &= P(\mathbf{w}|\mathbf{l})P(\mathbf{l}) \\ P(\mathbf{l}) &= \left(\frac{1}{t}\right)^n \\ P(\mathbf{w}|\mathbf{l}) &\sim unif. \end{aligned} \quad (3\text{–}5)$$

Although sparse and strong priors characterize people's prior beliefs better than uninformative priors (Lu, Yuille, et al., 2008), the adoption of uninformative priors allows us to examine the importance of belief distribution by comparing the Bayesian model to associative models that do not incorporate prior beliefs about causal strength. From Bayes's theorem, the posterior distribution of the links and weights can be calculated as

$$P(\mathbf{w}, \mathbf{l}|D) = \frac{1}{Z} P(D|\mathbf{w}, \mathbf{l})P(\mathbf{w}|\mathbf{l})P(\mathbf{l}). \quad (6)$$

The variable Z represents a normalizing constant. Given this joint probability distribution, we can answer questions about both causal structure (i.e., whether a cue prevents, causes, or does not influence the effect) and causal strength (i.e., how strongly the cue causes or prevents the effect). For causal structure, the posterior probability that a cue is a generative cause is the sum of the posterior probabilities of the causal graphs where the cue is generative. Accordingly, the posterior probability that cue i is causal is

$$P(l_i = 1|D) = \sum_{\mathbf{l}\,:\,l_i=1} \int P(\mathbf{w}, \mathbf{l}|D)d\mathbf{w}. \quad (7)$$

Note that Equation 7 marginalizes over the other causal links and over the weights. Analogous calculations apply for the posterior probabilities that cue i is preventive ($l_i = -1$) or noncausal ($l_i = 0$). To predict people's answers to causal structure questions, we use the mean value of l_i.

For causal strength questions, we define the mean causal strength of cue i as

$$\overline{w_i l_i} = \sum_{\mathbf{l}} \int w_i l_i P(\mathbf{w}, \mathbf{l}|D)d\mathbf{w}. \quad (8)$$

The mean causal strength ranges from –1.0 for a deterministic preventer to 1.0 for a deterministic generative cause.

Although the computations in these equations involve integration, the analytic integration of these equations is tractable for small data sets—like those in the present article—with the assistance of a computer algebra program.

This model explains the inferential dependencies in recovery from overshadowing and backward blocking by how belief distribution across the multiple explanations changes as disambiguating evidence emerges. As shown in Figure 1, when provided with ambiguous AB+ evidence, the model distributes belief across the explanations where at least one of the cues causes the effect. Subsequent evidence that Cue A does not cause the effect (A– trials in recovery from overshadowing) therefore implies that Cue B must. In contrast, subsequent evidence that Cue A causes the effect (A+ trials in backward blocking) leaves Cue B as a potential, if less probable, cause of the effect. The Bayesian model therefore predicts that people will be more certain about the causal influence of Cue B in recovery from overshadowing than in backward blocking.

Associative Models

In response to the failure of the R-W model (Rescorla & Wagner, 1972) to account for inferential dependencies, three major associative models have been developed: Van Hamme and Wasserman's (1994) learning rule, the comparator hypothesis (Denniston, Savastano, & Miller, 2001; Miller & Matzel, 1988; Stout & Miller, 2007), and the modified sometimes-opponent-process (SOP) model (Dickinson & Burke, 1996). All of these models assume that the learner learns and utilizes within-compound associations, but they differ in significant ways in implementing these computational constraints, and therefore yield distinct predictions for various paradigms. To illustrate the associative approach, we review the R-W model and Van Hamme and Wasserman's learning rule presently. The comparator hypothesis and the modified SOP model are presented in Appendix A.

Rescorla-Wagner model. The R-W model (Rescorla & Wagner, 1972) is the most well-known associative model. The R-W model adopts the following learning rule, which modifies the associations between a cue i and the effect e in order to reduce prediction error:

$$\Delta V_i = s_i s_e \left(T - \sum_j V_j\right). \quad (9)$$

In this learning rule, s_i and s_e are learning rate parameters, where s_i represents the salience of cue i when it is present ($s_i = \alpha$) or absent ($s_i = 0$), s_e represents the salience of e when it is present ($s_e = \beta_1$) or absent ($s_e = \beta_2$, where β_2 is typically assumed to be

a positive number less than β_1). T represents the actual presence ($T = 1$) or absence ($T = 0$) of e, and V_i represents the current association between cue i and e. The summation, which occurs over all cues present on a given trial, is the predicted strength of e. The difference between T and the summation therefore represents the prediction error (i.e., the observed value minus the expected value), and the model modifies the association between the cue and e so that the error will be smaller on the same trials in the future.

The R-W model accounts for many notable experimental findings, including forward blocking (A+ AB+; e.g., Kamin, 1969). Compared with a control condition without the initial A+ trials (i.e., AB+ alone), forward blocking produces a weaker association between cue B and e. The R-W model explains this finding, because it learns a strong association between A and e during the A+ trials. This stronger association leads to a smaller prediction error on the AB+ trials, leaving less room to increase the association between Cue B and e.

The explanation of inferential dependencies is more problematic, however. Consider the predictions of the R-W model when presented with recovery from overshadowing (AB+ A–) and backward blocking (AB+ A+). Because the R-W model does not modify the associations of absent cues ($s_i = 0$ for absent cues), it does not predict any learning about Cue B during the A+ or A– trials.

As mentioned earlier, other associative models have been proposed to explain the existence of inferential dependencies by linking the presented and nonpresented cues through within-compound associations. Van Hamme and Wasserman's (1994) learning rule is one such model.

Van Hamme and Wasserman's learning rule. Van Hamme and Wasserman (1994) modified the R-W model by positing (a) within-compound associations between the cues that are presented together and (b) a negative learning rate for *absent but expected* cues, for which the expectation comes from the established within-compound associations with a present cue. Van Hamme and Wasserman did not specify a formal mechanism that controls the formation of within-compound associations, but it is typically assumed that within-compound associations form between cues that are presented simultaneously. We adopt this assumption when deriving the predictions of their learning rule. The learning rule retains Equation 9 but assigns a negative learning rate to nonpresented cues that are expected on the basis of a within-compound association (i.e., expected but absent cues). More specifically, the learning rate of the cue is set to different values depending on whether the cue is present ($s_i = \alpha_1$), expected-but-absent ($s_i = \alpha_2$, where α_2 is negative), or unexpected-and-absent ($s_i = 0$). Like the R-W model, Van Hamme and Wasserman's learning rule allows the learning rate to vary as a function of the presence and absence of e.

These modifications allow the learning rule to explain the existence of inferential dependencies in recovery from overshadowing (AB+ A–) and backward blocking (AB+ A+). Figure 2 shows the asymptotic associations predicted by the model when $\alpha_2 = -\alpha_1$ and $\beta_1 = \beta_2$. On the AB+ trials, the learning rule infers—just as the R-W learning rule would—that e is explained by both co-occurring cues. The within-compound association posited by the modified learning rule, however, provides the basis for assigning Cue B a negative learning rate during subsequent A– or A+ trials. As a consequence, Van Hamme and Wasserman's (1994) learning rule predicts that the Cue B–effect association will decrease during the A+ trials of backward blocking and increase during the A– trials of recovery from overshadowing.

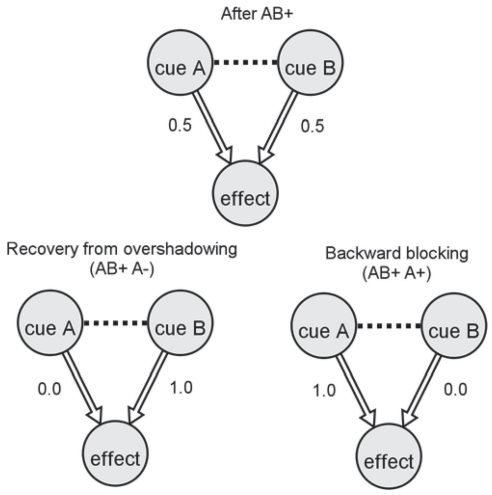

Figure 2. The explanation of recovery from overshadowing and backward blocking according to Van Hamme and Wasserman's (1994) learning rule. The arrows represent associations between the cues and the effect, and the adjacent numbers represent the asymptotic associative strengths. The dashed line represents a within-compound association between Cues A and B. The within-compound association supports learning about the association between Cue B and the effect on the A– and A+ trials, even though Cue B is not present on those trials.

The learning rule predicts that V_B, the association between Cue B and the effect, will be at least as close to 0.0 in backward blocking as it will be close to 1.0 in recovery from overshadowing. To see why, note that V_B will only asymptotically approach 0.5 on the AB+ trials and that A+ trials will lead to more learning than A– trials, given the standard assumption that the occurrence of the effect is more salient than its absence (i.e., $\beta_1 > \beta_2$). Thus, V_B will be closer to 0.0 in backward blocking than it is close to 1.0 in recovery from overshadowing. For a binary effect, associative strength can be viewed as the expected probability of the effect given the cue. This expected probability should presumably depend on both the reasoner's certainty that the cue causes the effect and the estimated causal strength of the cue assuming it causes the effect. Associative models do not differentiate between certainty and causal strength, however, so we treat associative strength as the associative estimate of certainty. Therefore, the learning rule predicts—in contrast to the Bayesian model—that people will be more certain that Cue B *does not cause* the effect in backward blocking than that the cue *causes* the effect in recovery from overshadowing (because V_B will be closer to zero in backward blocking than close to one in recovery from overshadowing).

Additional Models

We primarily aim to contrast belief-distribution models with associative models that adopt within-compound associations with regard to their representations of ambiguous evidence. But these models also make different assumptions about how causes combine. Our belief-distribution model posits that causal strengths

combine according to noisy-logical functions (see Equation 1), and the associative models posit that associations combine additively (e.g., the summation in Equation 9).

To explore the independent contributions of these differences, we consider two additional models: a linear Bayesian model with an additive combination rule and an associative model with a noisy-logical combination rule. Other researchers have considered Bayesian models with additive combination rules, either in isolation (e.g., Dayan & Kakade, 2001) or in comparison to noisy-logical combination rules (Griffiths & Tenenbaum, 2005; Kruschke, 2008). The linear Bayesian model that we develop, similar to these models, is identical to the noisy-logical Bayesian model presented earlier except that Equation 1 is replaced by the following equation:

$$P(e = 1|\mathbf{c}, \mathbf{l}, \mathbf{w}) = \sum_{g \in G} (w_g)^{c_g} - \sum_{p \in P} (w_p)^{c_p}. \quad (10)$$

Note that Equation 10 only produces a valid probability distribution when the weights are constrained so that the function never produces a value less than 0.0 or greater than 1.0. We constrain the weights as needed to achieve this goal. For example, when Cues A and B are presented together and assumed to be causal, we add a constraint that $w_a + w_b \leq 1.0$. We refer to the model using Equation 11 as the linear Bayesian model. The linear Bayesian model is closely related to models of causal learning based on the Kalman filter (e.g., Kruschke, 2008). We adopt the linear Bayesian model rather than the Kalman filter because the Kalman filter introduces additional assumptions that complicate the comparison to the noisy-logical Bayesian model. Note, for example, that the Kalman filter assumes that the power of a cause to produce its effect gradually changes over time. When we refer to "the Bayesian model" without qualification, we are referring to the belief-distribution model that uses the noisy-logical combinations rules in Equation 1.

The noisy-logical associative model, adapted from Danks, Griffiths, and Tenenbaum (2003), is identical to Van Hamme and Wasserman's (1994) learning rule except that Equation 9 is replaced with the following equation, which reflects the noisy-logical combination rule:

$$\Delta V_i = s_i s_e \left[T - \left[1 - \prod_{g:V_g>0} (1 - V_g) \right] \left[\prod_{p:V_p<0} (1 - |V_p|) \right] \right] \quad (11)$$

We assume that the products in this equation are only computed over the cues that are present on a given trial. Equation 11 performs error-correction in the same way as Equation 9 but derives predictions using the noisy-logical function rather than a linear function.

Distinguishing Belief Distribution and Within-Compound Associations

As illustrated earlier, belief-distribution models and associative models augmented with within-compound associations make different predictions regarding recovery from overshadowing and backward blocking. Experiments 1A and 1B tested some situations where the models predict different inferential dependencies.

In other situations, the associative models predict inferential dependencies, even though the belief-distribution model does not. Because the associative models predict the formation of within-compound associations whenever two cues are presented simultaneously, these models predict inferential dependencies even in the absence of ambiguity about the causal influence of the target cue. Consider a situation where the effect follows the presentation of one cue (A+) as well as the presentation of that cue and another cue (AB+). Now, suppose also that the effect is later observed to follow the presentation of the other cue (B+). Because associative models predict that the AB+ observations create a within-compound association between the cues, they predict that subsequent learning about Cue B, even though the causal status of Cue A is already known, might still influence inferences about Cue A. For example, Van Hamme and Wasserman's (1994) learning rule predicts that the B+ trials will diminish the Cue A–effect association substantially. Models that distribute belief across multiple explanations make the more intuitive prediction that *once there is no uncertainty regarding the causal influence of a cue*, beliefs about the cue will remain unchanged so long as new information does not contradict those beliefs. We tested these predictions in Experiment 2.

Experiments 1A and 1B

The different models often make competing predictions about the exact form of the inferential dependencies. For example, as mentioned previously, the Bayesian model predicts that people will be more certain that Cue B causes the effect in recovery from overshadowing than certain that Cue B does not cause the effect in backward blocking. Van Hamme and Wasserman's (1994) learning rule predicts the opposite: that people will be *less certain* that Cue B causes the effect in recovery from overshadowing than they will be certain that Cue B does not cause the effect in backward blocking. We tested these predictions in Experiment 1A.

For Experiment 1A, the Bayesian model, the modified SOP model, and the comparator hypothesis make similar predictions. We therefore consider some additional procedures in Experiment 1B. In particular, we investigate recovery from preventive overshadowing (A+ and ABC– trials followed by AB+ trials) and preventive backward blocking (A+ trials and ABC– trials followed by AB– trials). For both procedures, we were especially interested in whether participants would make inferences about Cue C during the AB trials. As becomes clear in the experimental results, these preventive procedures are useful for discriminating between the Bayesian and the other associative models.

Method

Participants. Thirty-two undergraduate students participated in Experiment 1A, and another 32 undergraduate students participated in Experiment 1B. All were from the University of California, Los Angeles and participated for course credit.

Materials and procedure. Participants were asked to diagnose the fruit allergies of the patients in a doctor's office by discovering the fruits that caused each patient's allergic reactions. The cover story explained that the diagnoses would be made by reviewing "fruit journals." Each fruit journal provided a daily log of the fruits that a patient ate and the occurrence of his or her allergic reactions. Participants in Experiment 1B also read that "fruit allergies can be both caused and prevented" in that "some

fruits may cause the allergic reaction and other fruits may prevent it."

Participants in each experiment viewed five fruit journals, whose contents are summarized in Table 1. The journals were presented in a randomized order. The same cue corresponded to different fruits across fruit journals (i.e., Cue A was a different fruit in different journals), and the assignment of the fruits to the cues in the journals was randomized. Each fruit journal contained multiple phases. The phases were presented sequentially, but the order of the trials within a phase was randomized. Each phase contained five trials of each trial type (e.g., there were five AB+ trials in the first phase of the recovery from overshadowing fruit journal).

Each trial displayed the pictures and names of whichever fruits the patient ate on that day. The fruits were displayed alone for 1.5 s, at which point a cartoon face appeared. The cartoon face signified whether the patient had an allergic reaction on that day: A smiley face with the text "ok" indicated that the patient did not have a reaction and a frowning face with the text "allergic reaction" indicated that the patient had a reaction. The trial ended after the fruits and cartoon face were displayed together for 2.0 s.

After a fruit journal was presented, we assessed the causal beliefs of the participants. In Experiment 1A, participants were asked whether the fruit caused or did nothing to influence the patient's allergic reactions. In Experiment 1B, participants were asked whether the fruit caused, prevented, or did nothing to influence the patient's allergic reactions. These questions assess causal structure (whether a causal relationship exists; Lu, Yuille, et al., 2008).

Responses were recorded on sliding scales with five tick marks in Experiment 1A and nine tick marks in Experiment 1B, with each mark labeled to encourage participants to distinguish between different degrees of a cue "maybe" causing an effect. In Experiments 1A and 1B, respectively, the leftmost mark (labeled "definitely not a cause") and middle mark (labeled "neither" cause nor preventer) corresponded to cues that did not influence the effect. The other marks corresponded to cues that were "possible (but not likely)," "somewhat likely," "probable," and "definite" causes and preventers. Responses were coded as integers ranging from 0 to 4 in Experiment 1A and −4 to 4 in Experiment 1B. Responses were divided by the highest possible response (4) to produce a causal rating with a maximum value of 1.0 (corresponding to a cue that "definitely" causes the effect).

Table 1
The Contents of the Fruit Journals Shown to Participants in Experiments 1A and 1B

Fruit journal	Phase 1	Phase 2	Phase 3
Experiment 1A			
RO	AB+ c+ d−	A−	
BB	AB+ c+ d−	A+	
RO control	AB+ c+ d−	d−	
BB control	AB+ c+ d−	c+	
No-trial control	AB+ c+ d−		
Experiment 1B			
pRO	A+ B− C− d−	A+ ABC−	A+ AB+
pBB	A+ B− C− d−	A+ ABC−	A+ AB−
pRO control	A+ B− C− d−	A+ ABC−	A+ Ad+
pBB control	A+ B− C− d−	A+ ABC−	A+ Ad−
No-trial control	A+ B− C− d−	A+ ABC−	A+

Note. RO = recovery from overshadowing; BB = backward blocking; pRO = recovery from preventive overshadowing; pBB = preventive backward blocking. Filler cues are written in lowercase.

Results

Table 2 shows the causal ratings for each cue. We report analyses of the causal ratings that provide the clearest tests of the models only: those of Cue B in Experiment 1A and Cue C in Experiment 1B. Figure 3 displays the causal ratings and model predictions for these cues. Observe that the causal ratings in the experimental and corresponding control conditions differed for recovery from overshadowing, $t(31) = 3.47$, $p < .01$; backward blocking, $t(31) = 2.18$, $p < .05$; recovery from preventive overshadowing, $t(31) = 3.74$, $p < .01$; and preventive backward blocking, $t(31) = 3.74$, $p < .01$. These differences reflect the existence of inferential dependencies. The causal ratings for the target cues did not differ significantly across the control conditions in either Experiment 1A, $F(2, 62) = 1.19$, $p = .31$, or Experiment 1B, $F(2, 62) = 2.14$, $p = .13$.

As expected, participants were more certain about the causal influence of the target cue in the recovery from overshadowing condition than in the backward blocking condition: Observe that the mean causal rating for Cue B in recovery from overshadowing (0.77) was much closer to 1.0 than the mean causal rating for Cue B in backward blocking (0.48) was close to zero. The difference in certainty was less pronounced between recovery from preventive overshadowing (−0.75 causal rating for Cue C) and preventive backward blocking (−0.29 causal rating for Cue C). To explore these trends in greater detail and analyze their statistical significance, we considered the proportion of participants who were certain about the influence of the target cue in each condition. In Experiment 1A, 15 out of 32 participants in the recovery from overshadowing condition believed that Cue B was a "definite" cause of the effect (i.e., provided a causal rating of 1.0), whereas none of the participants in the backward blocking condition indicated that Cue B was "definitely not a cause" (i.e., provided a causal rating of 0.0), $\chi^2(1, N = 15) = 13.07$, $p < .001$ (by McNemar's test). In Experiment 1B, 18 out of 32 participants in the recovery from preventive overshadowing condition believed that Cue C was a "definite" preventer (i.e., gave a rating of −1.0), whereas only eight out of 32 participants in the preventive backward blocking condition believed that Cue C was neither a cause nor a preventer (i.e., gave a rating of 0.0), $\chi^2(1, N = 18) = 4.5$, $p < .05$ (by McNemar's test).

Model predictions. The predictions of the associative models are parameter-dependent, so we set the parameters of the associative models to maximize the fit to the experimental results. Appendix B describes the procedure used to select the model parameters.

The Bayesian model correctly predicts that while the target cue in Experiment 1A was unambiguously causal in recovery from overshadowing ($l_i \approx 1.0$; see the leftmost bar of the Bayesian model predictions in Figure 3), its causal influence was ambiguous in backward blocking ($l_i \approx 0.5$; see the second bar). The Bayesian model also predicts the analogous findings for the preventive variants of these paradigms (Experiment 1B). None of the other models predict all four of these basic findings. The R-W model

Table 2
Causal Ratings for Each Cue in Experiments 1A and 1B

	Cue					
	A		B		C	
Fruit journal	*M*	*SD*	*M*	*SD*	*M*	*SD*
Experiment 1A						
RO (AB+ A−)	.10	.15	**.77**	.29		
BB (AB+ A+)	.90	.21	**.48**	.19		
RO control (AB+ d−)	.57	.17	**.59**	.14		
BB control (AB+ c+)	.56	.16	**.55**	.16		
No-trial control (AB+)	.59	.15	**.58**	.15		
Experiment 1B						
pRO (A+ ABC− AB+)	.94	.11	−.02	.14	**−.75**	.33
pBB (A+ ABC− AB−)	.94	.11	−.87	.23	**−.29**	.24
pRO control (A+ ABC− Ad+)	.88	.36	−.59	.23	**−.56**	.25
pBB control (A+ ABC− Ad−)	.93	.13	−.51	.22	**−.48**	.24
No-trial control (A+ ABC−)	.94	.11	−.48	.27	**−.51**	.26

Note. RO = recovery from overshadowing; BB = backward blocking; pRO = recovery from preventive overshadowing; pBB = preventive backward blocking. Filler cues are written in lowercase. The mean causal ratings for the critical cues (Cue B in Experiment 1A and Cue C in Experiment 1B) are displayed in bold.

simply fails to predict any retrospective revaluation. Van Hamme and Wasserman's (1994) rule erroneously predicts that the causal rating for the target cue will be slightly closer to 0.5 in recovery from overshadowing than in backward blocking. (The learning rule also predicts that the target cue is always close to −0.5 in Experiment 1B, although this prediction is parameter dependent.[2]) The comparator hypothesis predicts the causal ratings in Experiment 1A but erroneously predicts that the causal rating for Cue C will be nearly 0.0 (corresponding to a cue that is definitely not causal) in the recovery from preventive overshadowing condition of Experiment 1B (note the near-zero value of the leftmost bar for Experiment 1B in Figure 3). The higher order comparison process accounts for this erroneous prediction (for details, see Stout & Miller, 2007). Finally, the modified SOP model barely predicts any retrospective revaluation at all given its best fitting parameters. Furthermore, even when the qualitative predictions of the modified SOP are considered, it fails to predict the results (see Appendix A).

The noisy-logical associative model correctly predicts that recovery from overshadowing is less ambiguous than backward blocking but erroneously predicts more negative causal ratings in the preventive backward blocking condition than in the relevant control condition. The predictions of the two Bayesian models are similar, suggesting that the results are generally consistent with both the noisy-logical and linear combination rules. Note, however, that the linear Bayesian model erroneously predicts that people should be quite certain that the target cue is noncausal in both generative and preventive backward blocking conditions.

Overall, the standard and linear Bayesian model produced the highest rank-order correlations with the data (r_s = .98, *MSE* = 0.026 and r_s = .98, *MSE* = 0.019, respectively)—higher than that of the R-W model (r_s = .95, *MSE* = 0.00118), Van Hamme and Wasserman's (1994) learning rule (r_s = .97, *MSE* = 0.0089), the comparator hypothesis (r_s = .91, *MSE* = 0.064), the modified SOP model (r_s = .91, *MSE* = n/a), and the noisy-logical associative model (r_s = .96; *MSE* = 0.010). Although the *MSE*s of some of the associative models were lower than the *MSE* of the Bayesian model, this is not surprising given that the comparison is between the parameter-free Bayesian model and associative models with free parameters that were selected to maximize the fit to the data.

Discussion

Why was the Bayesian model more successful in explaining the inferential dependencies than the associative models? By the incorporation of deductive inference, the Bayesian model implicitly encodes not only that an inferential dependency exists but also its specific form. For example, if belief is distributed across those explanations where at least one of the cues causes the effect, it implies that (a) if one of the cues does not cause the effect, then the other *must* and (b) if one of the cues causes the effect, then the other *might*. Within-compound associations, on the other hand, do not use deductive logic and therefore do not directly encode the exact form of an inferential dependency. Instead, the form of the inferential dependency depends on interactions between the within-compound associations (which encode that an inferential dependency exists) and other mechanisms (which control how the inferential dependency is expressed). Experiments 1A and 1B illustrate that these interactions predict the correct inferential dependencies in some situations but not in others. Still, although none of the associative models predict the observed results and although belief distribution involves a more principled response to ambiguous evidence, some modification of one of the associative models might explain the results. We revisit this issue in the General Discussion, preferring to discuss the other implications of the results presently.

The belief-distribution approach implies that people will be able to flexibly encode inferential dependencies with different forms. For example, while AB+ trials in backward blocking usually lead people to infer that *at least* one of the cues causes the effect, there may be situations where AB+ evidence leads people to infer that *at most* one of the cues causes the effect. Consider a field biologist who identifies two new bird species on a remote island and observes that something on the island is splitting coconuts apart (AB+). Knowing that ecological niches are usually occupied by a single species, the biologist may believe that *at most* one of the bird species evolved to split coconuts. This belief distribution creates an inferential dependency that differs from the one typically observed in backward blocking: After learning that one bird species splits coconuts (A+), the biologist would be *certain* that the other species does not split coconuts. Mitchell, Killedar, and Lovibond (2005) manipulated whether participants' prior beliefs supported belief distributions like this one, and people produced different inferential dependencies for the different belief distribu-

[2] The learning rule predicts only small changes in belief during the final phase in Experiment 1B because retrospective revaluation of Cue C on the A and AB trials will proceed in opposite directions. For example, the preventive backward blocking condition alternated A+ and AB– trials. On the ABC– trials in the previous phase, the learning rule infers that $V_a \approx 1.0$, $V_b = V_c \approx -0.5$. As a consequence, the learning rule overpredicts the absent effect on the AB– trials. This should lead the learning rule to (among other things) decrease V_a. The decrease in V_a, however, would lead it to underpredict the effect on subsequent A+ trials. The net result is that V_c increases on AB– trials and decreases on A+ trials. The learning rule predicts the observed results more closely if retrospective revaluation of Cue C is assumed to occur on AB trials but not on A trials.

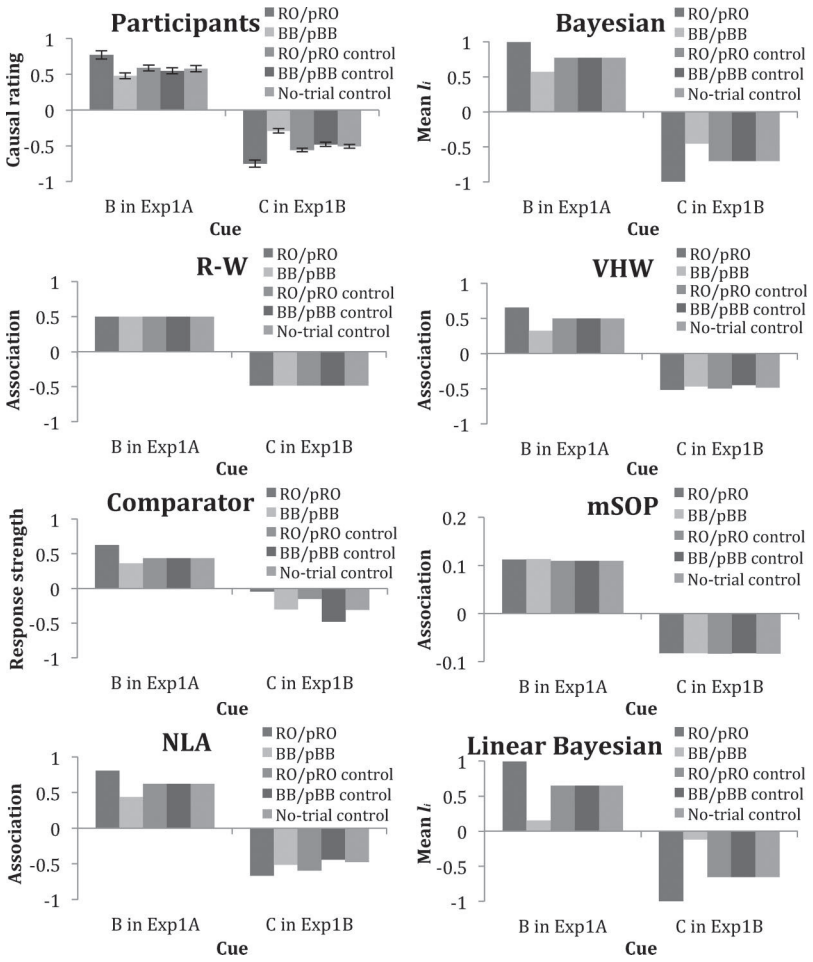

Figure 3. Causal ratings and model predictions for Cue B in Experiment 1A (B in Exp1A) and Cue C in Experiment 1B (C in Exp1B) as a function of experimental condition. Error bars indicate standard errors. R-W = Rescorla and Wagner (1972) model; VHW = Van Hamme and Wasserman's (1994) learning rule; mSOP = modified sometimes-opponent-process model; RO = recovery from overshadowing; BB = backward blocking; pRO = recovery from preventive overshadowing; pBB = preventive backward blocking; NLA = noisy-logical associative model.

tions. Additionally, the flexibility of the belief-distribution approach explains why different assumptions about how causes combine can support different inferential dependencies (e.g., Beckers, De Houwer, Pineno, & Miller, 2005; Lu, Rojas, et al., 2008). The associative models might explain these findings by adjusting the model parameters (e.g., by adjusting the sign of α_2 in Van Hamme and Wasserman's, 1994, learning rule) across the different tasks, but the rationale for these parameter adjustments is unclear.

It may be surprising that our experiment found evidence for backward blocking (and its preventive variant), given that recovery from overshadowing is often observed in situations where backward blocking is not (Corlett et al., 2004; Larkin et al., 1998; see also Beckers, De Houwer, Pineno, & Miller, 2005; Lovibond, Been, Mitchell, Bouton, & Frohardt, 2003; Vandorpe & De Houwer, 2005). How are we to reconcile our finding (and similar findings such as Wasserman & Berglan, 1998; Wasserman & Castro, 2005) with these other findings? The Bayesian model offers one possible explanation. Although the Bayesian model predicts backward blocking, it also predicts that target cue will remain ambiguous in backward blocking: note that while the model predicts that $P(B \rightarrow E)$ is lower after backward blocking than after the initial evidence (see Figure 1), $P(B \rightarrow E)$ does not approach 1.0 or 0.0 in either situation. To the extent that people fail to distinguish between different degrees of the target cue "maybe" causing the effect, we would expect them to provide the same causal ratings for the target cue in backward blocking as they provide in relevant control conditions (e.g., AB+). In short, some experimental measures of causal beliefs may not be sensitive enough to detect all of the changes in the participants' beliefs. Indeed, experiments demonstrating backward blocking have sometimes used more sensitive measures than experiments that failed to demonstrate backward blocking. For example, Wasserman and Berglan (1998) labeled the causal rating scale so that some of its tick marks corresponded to cues that "definitely would not," "probably would not," "possibly," "probably would," or "defi-

nitely would" cause the effect. Our experiment also used a detailed rating scale. The experiments that did not show backward blocking have used less detailed rating scales, labeling only the tick marks corresponding to a cue that "definitely would not," "possibly," or "definitely would" cause the effect (e.g., Larkin et al., 1998; Lovibond et al., 2003). Furthermore, when we replicated Experiment 1B with a less detailed rating scale, there was no evidence of preventive backward blocking (Carroll, Cheng, & Lu, 2010). Of course, the specificity of the rating scale is not the only factor that will influence whether backward blocking is observed (e.g., see Beckers et al., 2005; De Houwer, Beckers, & Glautier, 2002; Lovibond et al., 2003; Miller & Matute, 1996).

Experiment 2

In Experiment 2, we investigated whether inferential dependencies form between any cues that are presented simultaneously, as the associative models predict. Table 3 shows the experimental design. The two-cause condition combines forward and backward blocking. The one-cause condition combines latent inhibition and backward blocking. The associative models predict that within-compound associations will form between Cues A and B during the AB+ trials in the two-cause condition and between Cues C and D during the CD+ trials in the one-cause condition. This leads to the anomalous prediction that causal beliefs about Cues A and C may be subject to revision as the participant learns about Cues B and D.

Moreover, all four associative models predict that in the one-cause condition the Cue C–effect association will increase during the second learning phase. To see why, consider the predictions of the R-W model. The R-W model predicts that Cue C–effect and Cue D–effect associations will be zero prior to the CD+ trials, so the R-W model will have a large prediction error on CD+ trials. In response to this prediction error, the model will increase the Cue C–effect and Cue D–effect associations. The other associative models predict increases in the Cue C–effect association for similar reasons. Belief-distribution models, in contrast, make a more intuitive prediction that Cue C will still be viewed as noncausal following the CD+ trials. Admittedly, variants of these associative models, by introducing a competitive context or by proposing a mechanism that reduces attention to familiar cues, predict the stability of the Cue C–effect associations. We argue in the Discussion section, however, that these modifications will ultimately prove unsatisfactory.

The competition control condition, which makes use of the recovery from overshadowing procedure, serves to confirm that our experimental procedure allows within-compound associations to form. We expected that an inferential dependency would form between Cues E and F on the EF+ trials and that the dependency would be revealed after the subsequent F– trials. (If the procedure does not lead to an inferential dependency, then the associative models could explain stable causal ratings for Cues A and C by setting the parameters to eliminate inferential dependencies; e.g., by setting $\alpha_2 = .0$ for Van Hamme and Wasserman's, 1994, learning rule.)

Table 3
Experimental Design of Experiment 2

Condition	Phase 1	Phase 2	Phase 3
Two causes	A+ w−	AB+	B+
One cause	C− x+	CD+	D+
Competition control	y+ z−	EF+	F−

Note. Lowercase letters represent filler cues.

Our experimental method recalls a series of experiments where Shanks and colleagues (Shanks, Charles, Darby, & Azmi, 1998; Shanks, Darby, & Charles, 1998) demonstrated that people's causal beliefs remain stable in certain situations where associative models predict otherwise. Shanks and colleagues explained their experimental findings by appealing to *configural processing*: processing where configurations of stimuli are represented as undivided entities. In configural models (e.g., Pearce, 1987, 1994), the stimulus composed of Cues X and Y together (XY) is represented independently of the stimuli composed of Cue X alone (X) and Cue Y alone (Y), and learning about the XY configuration can exert an influence on the predicted outcome separately from its constituent elements. Because outcome prediction is dependent on generalization due to similarity among stimuli (e.g., Stimuli XY and X both have Cue X in common), one might think that configural models can explain inferential dependencies without the instability of the models positing within-compound associations. However, the generalization in configural models cannot explain backward blocking or recovery from overshadowing. For example, although X+ trials following XY+ trials would increase responding to the XY stimulus, they would not influence the strength of responding to the Y stimulus (e.g., Pearce, 1987, 1994). The reasons for this are that (a) because Stimuli X and Y are dissimilar, there is no direct generalization from X to Y and (b) because configural models (e.g., Pearce, 1987, 1994) typically assume that the X+ trials alter responding to the XY stimulus without altering the XY–effect association per se, there is no indirect generalization. (Even if the models supported direct or indirect generalization between X and Y, the generalization would be in the same rather than the opposite direction as the "competing" cue, and the model would erroneously predict that responding to Cue Y would *increase* following the X+ trials.) Moreover, in our experiments, unlike Shanks et al.'s, because the individual cues can already explain the data, there is no motivation for configural cues.

Method

Participants. Eleven undergraduate students at the University of California, Los Angeles participated for course credit.

Materials and procedure. Except where noted, the materials and procedure were identical to those in the previous experiments. Participants viewed the fruit journals shown in Table 3, and there were four trials of each trial type.

To measure how causal beliefs changed over the course of the experiment, we assessed the causal beliefs of the participants after each phase in each fruit journal. For each fruit presented in the fruit journal up to that point, participants were asked whether the patient would have an allergic reaction on a day when he or she ate the fruit (a causal strength question, Lu, Yuille, et al., 2008). Responses were made on a sliding scale with seven tick marks, with the leftmost mark labeled "definitely not," the middle mark labeled "maybe," and rightmost mark labeled "definitely." No other tick marks were labeled. Responses were coded as integers ranging

from 0 (*definitely not*) to 6 (*definitely*) and then divided by the highest possible response (6) to produce a causal rating that ranged from 0.0 to 1.0.

Results

Table 4 lists the causal ratings given by the participants, and Figure 4 shows the causal ratings and model predictions for the cues that are most relevant for differentiating between the models. Observe that the causal ratings for Cues A and C were very stable over the course of the experiment. In fact, each participant gave identical causal ratings for these cues in the second and third learning phases. By comparison, the causal rating for Cue E clearly changed after participants learned that Cue F did not cause the effect (F–). Moreover, contrary to the predictions of the associative models, none of the participants gave a higher causal rating for Cue C in the second learning phase than in the first learning phase. Statistical tests confirmed that the causal ratings for Cue E changed across the phases, $t(10) = 5.04$, $p < .001$, and that the causal ratings for Cues A and C were not significantly different across the three phases. Every participant gave the same causal rating for Cue A in each phase; only a few participants gave Cue C a different rating across any of the phases, and the mean differences in the causal ratings for Cue C between Phases 1 and 2 were small, in the opposite direction as predicted by all of the associative models, and nonsignificant, $F(2, 20) = 1.00$, $p = .39$.[3]

Model predictions. The relative stabilities of the causal ratings for Cues A, C, and E were only predicted by the Bayesian model. Consistent with our findings in Experiments 1A and 1B, the success of the Bayesian model depended on both its ability to distribute belief across multiple explanations and on its assumptions about how causes combine, as revealed by the better fit of the Bayesian model compared to the noisy-logical associative model (which predicts substantial instability for the causal ratings for Cues A and C) and the linear Bayesian model (which predicts substantial instability for the causal ratings for Cue A). We defer a discussion about why the linear Bayesian model predicts changes in the causal ratings for Cue A until the Discussion section.

The R-W model cannot explain why the causal ratings for Cue E changed. None of the other associative models can explain why the causal ratings for Cues A and C remain stable. (As before, we selected the parameters of the associative models to maximize their fit to the results. See Appendix B.) First, all the associative models erroneously predict that the causal ratings for Cue C will increase substantially during the second learning phase. In addition, Van Hamme and Wasserman's (1994) learning rule erroneously predicts that people become much less certain that Cues A and C cause the effect after the third learning phase. In fact, because the prediction errors of the learning rule will be much larger on the B+ trials than the F– trials (given that A+ trials precede the AB+ trials, $V_B \approx 0.0$ after the AB+ trials; on the other hand, following the EF+ trials, $V_F \approx 0.5$), the learning rule predicts that the causal ratings for Cue A will change *more than* the causal ratings for Cue E. The comparator hypothesis and the modified SOP model also erroneously predict that within-compound associations formed during Phase 1 will lead to considerable instability in the causal ratings for Cues A and C in subsequent phases. These predictions are discussed in greater detail in Appendix A.

The Bayesian model based on belief distribution clearly showed better fit with human performance than all other models. Besides explaining the qualitative pattern of results well, it provided a better overall fit ($r_s = .91$; $MSE = 0.0074$) to the data than the R-W model ($r_s = .78$; $MSE = 0.067$), Van Hamme and Wasserman's (1994) learning rule ($r_s = .80$; $MSE = 0.060$), the comparator hypothesis ($r_s = .74$; $MSE = 0.058$), the modified SOP model ($r_s = .80$; $MSE =$ n/a), the linear Bayesian model ($r_s = .87$; $MSE = 0.043$), and the noisy-logical associative model ($r_s = .76$; $MSE = 0.062$).

Table 4
Causal Ratings for Each Cue in Experiment 2

	Causal rating					
	Phase 1		Phase 2		Phase 3	
Cue	*M*	*SD*	*M*	*SD*	*M*	*SD*
A	**.97**	**.07**	**.97**	**.07**	**.97**	**.07**
B	—		.50	.00	.94	.15
C	**.15**	**.28**	**.06**	**.15**	**.06**	**.15**
D	—		.94	.15	.94	.15
E	—		**.52**	**.05**	**.86**	**.23**
F	—		.52	.05	.12	.30

Note. The causal ratings for the critical cues are displayed in bold. A dash indicates that participants did not provide causal ratings for the given cue in the given phase. (Participants were not asked to provide causal ratings for yet-to-be-encountered cues.)

Discussion

Contrary to the predictions of the associative models, some simultaneous presentations of multiple cues simply do not create inferential dependencies between those cues. Instead, as predicted by the Bayesian model, inferential dependencies typically do not form between cues with unambiguous causal influences (e.g., Cues A and C). Furthermore, given that participants learned an inferential dependency between Cues E and F, the results cannot be explained by configural processing (e.g., Shanks, Charles, et al., 1998; Shanks, Darby, & Charles, 1998) or by the impairment of the processes that form and utilize within-compound associations.

The specific associative models considered here cannot explain the stability of causal estimates regarding Cue C in the second learning phase. However, some associative accounts—including the comparator hypothesis when given a representation of the context—explain latent inhibition (Lubow & Moore, 1959), where X+ trials produce a weaker cue–effect association when they are preceded by X– trials. Might some associative accounts therefore explain the present results? There are reasons to believe otherwise. Some of the associative models

[3] Because there was a missing data cell (Cue E in the first phase), we did not perform a statistical test on the Cue × Phase interaction across the three phases. A quick glance at the results, however, should confirm that the interaction exists. Furthermore, an ANOVA performed on the causal ratings for Cues A, C, and E in the second and third phases revealed a significant Cue × Phase interaction, $F(2, 20) = 25.4$, $p < .001$.

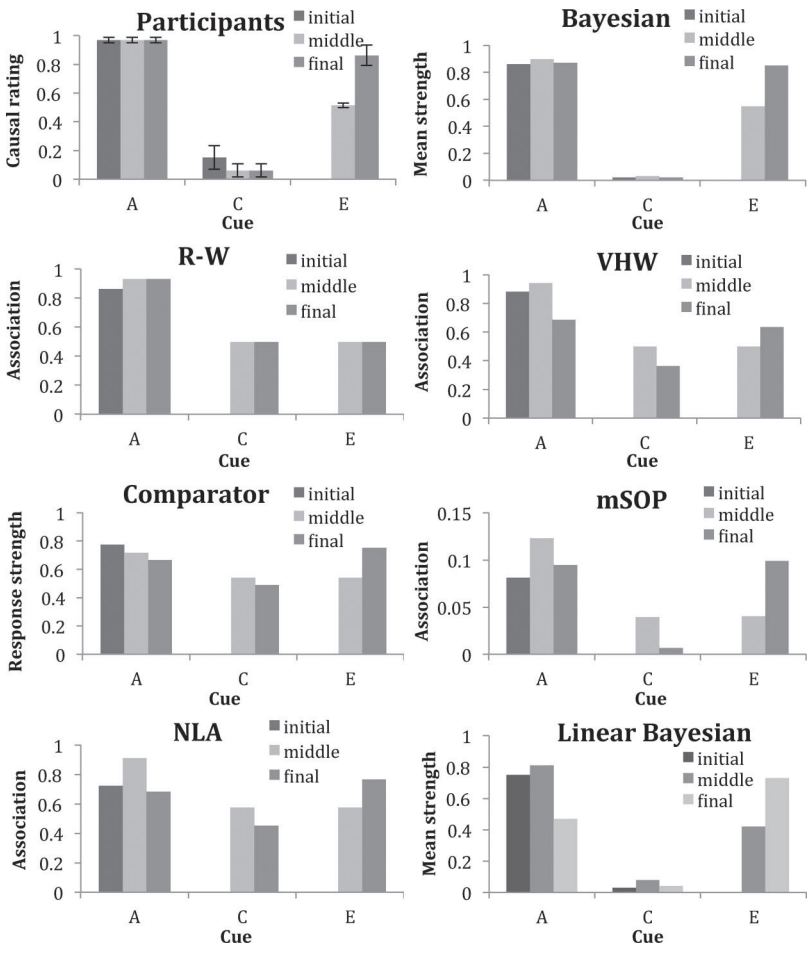

Figure 4. Causal ratings and model predictions for selected cues in Experiment 2. Note that only the Bayesian model predicts the relative stabilities of the causal ratings for the cues. Error bars correspond to standard errors. R-W = Rescorla and Wagner (1972) model; VHW = Van Hamme and Wasserman's (1994) learning rule; mSOP = modified sometimes-opponent-process model; NLA = noisy-logical associative model.

explain latent inhibition by positing that the context is established as a competitor of Cue X during the X– trials. While a competitive context might retard the increase in the causal ratings for Cue C, it cannot easily explain the complete absence of such an increase (in fact, none of the participants gave higher causal ratings for Cue C in the second learning phase). Moreover, we note that the comparator hypothesis predicts that the context would be a weaker competitor in the present experimental procedure than in latent inhibition (for reasons having to do with the higher order comparison process; for details, see Blaisdell, Bristol, Gunther, & Miller, 1998). Other associative models explain latent inhibition by positing that people pay less attention to familiar cues (e.g., McLaren & Mackintosh, 2000; Pearce & Hall, 1980). While a mechanism that dramatically reduces attention to familiar cues would allow an associative model to explain the stability of the causal ratings for Cue C, it would also incorrectly predict that the causal ratings for Cue E (a cue that would be familiar after the EF+ trials) would remain unchanged during the F– trials.

While we have argued that configural processing cannot explain our experimental results, configural processing or conjunctive causation may be important in other situations. The data from Experiment 2 can be explained by "simple" causes combining in accordance with the noisy-logical combination rules, but causes are not always simple (e.g., Novick & Cheng, 2004; Shanks, Charles, et al., 1998; Shanks, Darby, & Charles, 1998). In situations where simple causes cannot explain the data, people would be more likely to invoke conjunctive causes or to rely on configural processing.

Readers may be surprised that, despite our suggestion that belief-distribution accounts do not predict inferential dependencies between cues with unambiguous causal influences unless the new observations contradict past beliefs, the linear Bayesian model predicts that the causal ratings for Cue A will change during the final learning phase. Our suggestion holds, because for the linear Bayesian model, the new observations do contradict past beliefs. To see why, consider the predictions of the linear Bayesian model regarding Cues A and B (Kruschke,

2008, offered a similar explanation in the context of deriving predictions for the Kalman filter). From the A+ and AB+ trials, the model infers that $w_a \approx 1.0$, $w_a + w_b \approx 1.0$, and $w_b \approx 0.0$. The subsequent B+ trials add an additional constraint that $w_b \approx 1.0$, but this constraint directly contradicts the previous constraint that $w_b \approx 0.0$. Given these incompatible constraints, the linear Bayesian model selects intermediate values of w_a and w_b, thereby predicting smaller causal ratings for Cue A following the B+ trials.

General Discussion

The belief-distribution and associative approaches have profoundly different implications for our conception of human causal representations. Under the associative approach, the reasoner is assumed to learn within-compound associations and maintain a single hypothesis about the causes of the effect. In contrast, the belief-distribution approach postulates that humans construct and maintain multiple hypotheses, coding the uncertainty associated with each. The latter approach, when instantiated with noisy-logical generating functions, implies that humans will exhibit greater flexibility and logical consistency in the use of new information to update their beliefs about multiple alternative possible explanations of the data. Our results show that belief-distribution accounts offer principled and parsimonious explanations for inferential dependencies and provide a better account of people's inferences than associative models. The associative models that we considered failed to explain the form of inferential dependencies (Experiments 1A and 1B) and predicted inferential dependencies in situations where they were not observed (Experiment 2). Their failure indicates that current associative models, which conflate all of the possible explanations of the evidence into a single network state, cannot capture the logical consistency and flexibility of human causal inference.

Because we only examined the predictions of two variants of a single belief-distribution model (the noisy-logical and linear Bayesian models) and four associative models, one might question the generality of our conclusions regarding belief distribution. While some caution is warranted, there are strong reasons to believe that our conclusions generalize beyond these specific models. First, there are belief-distribution models that explain inferential dependencies without invoking probability (e.g., propositional models of causal inference; De Houwer et al., 2005; Lovibond, 2003; Mitchell, De Houwer, & Lovibond, 2009). Propositional models have been applied to explain backward blocking and recovery from overshadowing, and they would explain our experimental results as well. There are reasons to believe that probabilistic models may provide a more robust account of everyday causal reasoning than propositional models (Oaksford & Chater, 2007), but propositional and probabilistic models make the same predictions in many circumstances, and our Bayesian model can be regarded as a rational quantitative extension of propositional models.

Second, there is a general case to be made for the inadequacy of within-compound associations. We know of no associative model that explains our experimental results, and there is no obvious modification that would allow within-compound associations to explain the results. For example, Experiment 2 demonstrated that within-compound associations can be problematic when they form between cues with known causal influences. Yet by what means could an associative model prevent this? Without belief distribution and a representation of ambiguity, associative models cannot track whether a cue's influence is known. Although other features of the cue may be highly correlated with knowledge of a cue's influence, these correlations are imperfect. For example, although familiarity and certainty are highly correlated (familiar cues tend to have known causal influences), familiar cues can be constantly confounded and have unknown causal influences.

Although we have stressed the consequences of failing to distribute belief across multiple explanations, causal models differ on many other dimensions. On some of these other dimensions, associative models offer better accounts of the experimental results than our Bayesian model (for one review, see Perales & Shanks, 2007). For instance, while associative models offer detailed explanations for the influence of trial order, surprise (e.g., Pearce & Hall, 1980), and cue salience, our Bayesian model cannot explain why any of these factors influence people's inferences. Additionally, there may be other experimental procedures that encourage associative processing to a greater extent than ours. Our experiments presented no more than a few cues to the participant at any given point and did not provide trial-by-trial feedback; associative processing may be more prominent in other circumstances. Clearly, to offer a complete explanation of causal reasoning, our Bayesian model would require extension.

Because Bayesian models can be viewed in some respects as extensions of associative models (Kruschke, 2008), other Bayesian models may be able to offer a more complete account of causal inference by incorporating the many insights produced by work on associative models. Indeed, other Bayesian models of causal learning are sensitive to trial order (e.g., Daw et al., 2008; Kruschke, 2006; Lu, Rojas, et al., 2008) and the unexpectedness of observations (Courville, Daw, & Touretzky, 2006). For example, it should be possible to extend the present model using the framework of sequential Bayesian inference to account for some aspects of dynamic causal learning (e.g., Lu, Rojas, et al., 2008).

Given that causal models differ on many dimensions, direct comparisons between the models is not always as informative as one would hope. We believe that a more promising research strategy involves identifying the dimensions of variation across models of causal inference and testing the role of these differences in causal inference. Associative models typically explain learning on a trial-by-trial basis by appealing to error correction, assume that associations are represented as punctate values, and assume that the strength of the effect is an additive function of the associative strengths of its causes. Bayesian models typically make inferences from summarized data, distribute belief across multiple parameter-values and explanations, and assume a mechanism that can incorporate prior beliefs. Propositional models distribute belief across multiple causal structures but do not distribute belief across multiple parameter values. Ideally, research should aim to determine whether causal learning is propositional or probabilistic, whether it requires belief distribution, whether it is penetrable to language and instruction, whether it requires a priori causal assumptions, and so on.

The present experiments contribute to this endeavor by illustrating the importance of representing ambiguity by distributing belief across multiple explanations. Given the prevalence of ambiguous evidence in everyday causal reasoning, a representation of

ambiguity will prove useful when reasoning about causal evidence. This is something that the noisy-logical Bayesian models of causal inference clearly have but that associative models and within-compound associations fail to approximate. Belief distribution—whether done through probabilistic inference, propositional reasoning, or otherwise—plays an important role in explaining how people reason about inferential dependencies and ambiguous evidence.

References

Aitken, M. R., & Dickinson, A. (2005). Simulations of a modified SOP model applied to retrospective revaluation of human causal learning. *Learning & Behavior, 33,* 147–159. doi:10.3758/BF03196059

Beckers, T., De Houwer, J., Pineno, O., & Miller, R. R. (2005). Outcome additivity and outcome maximality influence cue competition in human causal learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 238–249. doi:10.1037/0278-7393.31.2.238

Blaisdell, A. P., Bristol, A. S., Gunther, L. M., & Miller, R. R. (1998). Overshadowing and latent inhibition counteract each other: Support for the comparator hypothesis. *Journal of Experimental Psychology: Animal Behavior Processes, 24,* 335–351. doi:10.1037/0097-7403.24.3.335

Carroll, C. D., Cheng, P. W., & Lu, H. (2010). Uncertainty in causal inference: The case of retrospective revaluation. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd annual conference of the Cognitive Science Society* (pp. 913–918). Austin, TX: Cognitive Science Society.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104,* 367–405. doi:10.1037/0033-295X.104.2.367

Corlett, P. R., Aitken, M. R., Dickinson, A., Shanks, D. R., Honey, G. D., Honey, R. A., . . . Fletcher, P. C. (2004). Prediction error during retrospective revaluation of causal associations in humans: fMRI evidence in favor of an associative model of learning. *Neuron, 44,* 877–888. doi:10.1016/S0896-6273(04)00756-1

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences, 10,* 295–300. doi:10.1016/j.tics.2006.05.004

Danks, D., Griffiths, T. L., & Tenenbaum, J. B. (2003). Dynamical causal learning. In S. Becker, S. Thrun, & K. Obermayer (Ed.), *Advances in neural information processing systems* 15 (pp. 67–74). Cambridge, MA: MIT Press.

Daw, N. D., Courville, A. C., & Dayan, P. (2008). Semi-rational models of conditioning: The case of trial order. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 427–448). New York, NY: Oxford University Press. doi:10.1093/acprof:oso/9780199216093.003.0019

Dayan, P., & Kakade, S. (2001). Explaining away in weight space. In T. Leen, T. Dietterich. & V. Tresp (Ed.), *Advances in neural information processing systems 13* (pp. 451–457). Cambridge, MA: MIT Press.

De Houwer, J., Beckers, T., & Glautier, S. (2002). Outcome and cue properties modulate blocking. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 55,* 965–985. doi:10.1080/02724980143000578

De Houwer, J., Beckers, T., & Vandorpe, S. (2005). Evidence for the role of higher order reasoning processes in cue competition and other learning phenomena. *Learning & Behavior, 33,* 239–249. doi:10.3758/BF03196066

Denniston, J. C., Savastano, H. I., & Miller, R. R. (2001). Learning by contiguity, responding by relative strength: The extended comparator hypothesis. In R. R. Mowrer & S. B. Klein (Eds.), *Handbook of contemporary learning theories* (pp. 65–117). Hillsdale, NJ: Erlbaum.

Dickinson, A., & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgments. *The Quarterly Journal of Experimental Psychology B: Comparative and Physiological Psychology, 49,* 60–80.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification*. New York, NY: Wiley.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51,* 334–384. doi:10.1016/j.cogpsych.2005.05.004

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review, 116,* 661–716. doi:10.1037/a0017201

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology, 62,* 135–163. doi:10.1146/annurev.psych.121208.131634

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511790423

Kamin, L. J. (1969). Selective association and conditioning. In N. J. Mackintosh & W. K. Honig (Eds.), *Selective association and conditioning* (pp. 42–64). Halifax, Nova Scotia, Canada: Dalhousie University Press.

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

Kruschke, J. K. (2006). Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychological Review, 113,* 677–699. doi:10.1037/0033-295X.113.4.677

Kruschke, J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning & Behavior, 36,* 210–226. doi:10.3758/LB.36.3.210

Larkin, M. J., Aitken, M. R., & Dickinson, A. (1998). Retrospective revaluation of causal judgments under positive and negative contingencies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1331–1352. doi:10.1037/0278-7393.24.6.1331

Lovibond, P. F. (2003). Causal beliefs and conditioned responses: Retrospective revaluation induced by experience and instruction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 97–106. doi:10.1037/0278-7393.29.1.97

Lovibond, P. F., Been, S.-l., Mitchell, C. J., Bouton, M. E., & Frohardt, R. (2003). Forward and backward blocking of causal judgment is enhanced by additivity of effect magnitude. *Memory & Cognition, 31,* 133–142. doi:10.3758/BF03196088

Lu, H., Rojas, R. R., Becker, T., & Yuille, A. (2008). Sequential causal learning in humans and rats. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th annual conference of the Cognitive Science Society* (pp. 64–70). Austin, TX: Cognitive Science Society.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review, 115,* 955–984. doi:10.1037/a0013256

Lubow, R. E., & Moore, A. U. (1959). Latent inhibition: The effect of nonreinforced exposure to the conditioned stimulus. *Journal of Comparative and Physiological Psychology, 52,* 415–419. doi:10.1037/h0046700

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. doi:10.1016/0023-9690(85)90023-2

McLaren, I. P., & Mackintosh, N. J. (2000). An elemental model of associative learning: I. Latent inhibition and perceptual learning. *Animal Learning & Behavior, 28,* 211–246. doi:10.3758/BF03200258

Miller, R. R., & Matute, H. (1996). Biological significance in forward and backward blocking: Resolution of a discrepancy between animal conditioning and human causal judgment. *Journal of Experimental Psychology: General, 125,* 370–386. doi:10.1037/0096-3445.125.4.370

Miller, R. R., & Matzel, L. D. (1988). The comparator hypothesis: A

response rule for the expression of associations. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. *2,* pp. 51–92). San Diego, CA: Academic Press.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences, 32,* 183–198. doi:10.1017/S0140525X09000855

Mitchell, C. J., Killedar, A., & Lovibond, P. F. (2005). Inference-based retrospective revaluation in human causal judgments requires knowledge of within-compound relationships. *Journal of Experimental Psychology: Animal Behavior Processes, 31,* 418–424. doi:10.1037/0097-7403.31.4.418

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review, 111,* 455–485. doi:10.1037/0033-295X.111.2.455

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford, England: Oxford University Press.

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review, 94,* 61–73. doi:10.1037/0033-295X.94.1.61

Pearce, J. M. (1994). Similarity and discrimination: A selective review and a connectionist model. *Psychological Review, 101,* 587–607. doi:10.1037/0033-295X.101.4.587

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review, 87,* 532–552. doi:10.1037/0033-295X.87.6.532

Perales, J. C., & Shanks, D. R. (2007). Models of covariation-based causal judgment: A review and synthesis. *Psychonomic Bulletin & Review, 14,* 577–596. doi:10.3758/BF03196807

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

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgement. *The Quarterly Journal of Experimental Psychology B: Comparative and Physiological Psychology, 37,* 1–21.

Shanks, D. R., Charles, D., Darby, R. J., & Azmi, A. (1998). Configural processes in human associative learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1353–1378. doi:10.1037/0278-7393.24.6.1353

Shanks, D. R., Darby, R. J., & Charles, D. (1998). Resistance to interference in human associative learning: Evidence of configural processing. *Journal of Experimental Psychology: Animal Behavior Processes, 24,* 136–150. doi:10.1037/0097-7403.24.2.136

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science, 28,* 303–333.

Stout, S. C., & Miller, R. R. (2007). Sometimes-competing cue retrieval (SOCR): A formalization of the comparator hypothesis. *Psychological Review, 114,* 759–783. doi:10.1037/0033-295X.114.3.759

Vandorpe, S., & De Houwer, J. (2005). A comparison of forward blocking and reduced overshadowing in human causal judgment. *Psychonomic Bulletin & Review, 12,* 945–949. doi:10.3758/BF03196790

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation, 25,* 127–151. doi:10.1006/lmot.1994.1008

Wagner, A. R. (1981). SOP: A model of automatic memory processing in animal behavior. In N. E. Spear & R. R. Miller (Eds.), *Information processing in animals: Memory mechanisms* (pp. 5–47). Hillsdale, NJ: Erlbaum.

Wasserman, E. A., & Berglan, L. R. (1998). Backward blocking and recovery from overshadowing in human causal judgment: The role of within-compound associations. *The Quarterly Journal of Experimental Psychology B: Comparative and Physiological Psychology, 51,* 121–138.

Wasserman, E. A., & Castro, L. (2005). Surprise and change: Variations in the strength of present and absent cues in causal learning. *Learning & Behavior, 33,* 131–146. doi:10.3758/BF03196058

(*Appendices follow*)

Appendix A

The Comparator Hypothesis and the Modified SOP Model

In this appendix, we describe the comparator hypothesis and the modified sometimes-opponent-process (SOP) model. We then discuss the predictions of the models in the experiments.

The Comparator Hypothesis

The comparator hypothesis (Denniston et al., 2001; Miller & Matzel, 1988; Stout & Miller, 2007) uses response strengths, as opposed to the typical cue–effect associations, to predict the effect. According to the comparator hypothesis, the response strength of a cue (the extent to which it leads to the expectation of the effect) is computed by comparing its *direct* and *indirect* activation of the effect. The cue's direct activation of the effect is the association between the cue and the effect, and the indirect activation of the effect is the product of the associations along an indirect path to the effect that traverses a within-compound association. A cue is viewed as causal to the extent that its direct activation of the effect exceeds its indirect activation of the effect. Although the comparator hypothesis also posits a more complicated higher order comparison process, this process rarely influences the predictions of the model in the present article. We note its influence when it is relevant. Interested readers can find the details of the higher order comparison process in Stout and Miller (2007).

The comparator hypothesis updates associations using a modification of the R-W learning rule (Stout & Miller, 2007):

$$\Delta V_{i,j} = s_i s_j (T - V_{i,j}). \quad \text{(A1)}$$

There are two important differences between Equation A1 and the R-W learning rule (Equation 9). First, the updated equation is applied to learn within-compound associations in addition to cue–effect associations: $V_{i,j}$ represents the association from cue i to the variable indexed by j. Depending on whether the variable indexed by j is a cue or the an effect, $V_{i,j}$ represents the strength of either a within-compound association or a cue–effect association. Second, Equation 10 calculates the prediction error relative to the prediction of a single association $V_{i,j}$, rather than relative to a sum of the associations of the present cues. The consequences of this modification can be seen by considering the AB+ trials. During these trials, the standard R-W rule predicts that the cue–effect associations will approach 0.5 (see Figure 2), but the comparator hypothesis predicts that the cue–effect associations will approach 1.0 (see Figure A1).

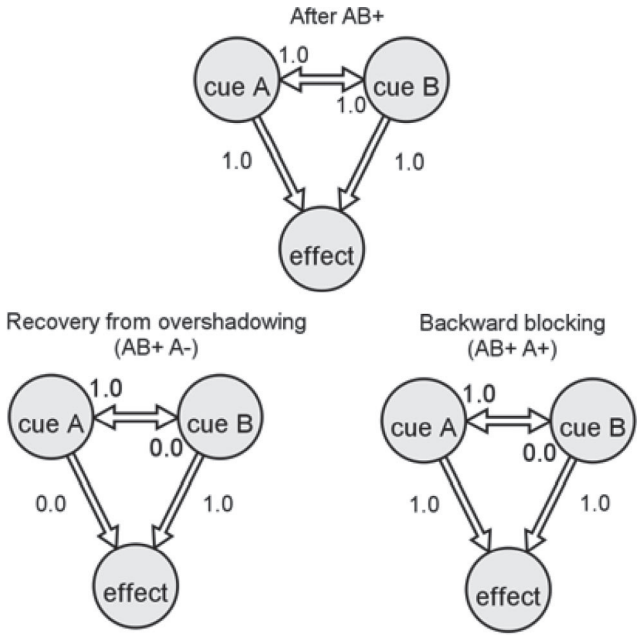

Figure A1. The asymptotic associations as predicted by the comparator hypothesis in response to recovery from overshadowing and backward blocking. The response to Cue B is calculated by comparing its direct activation of the effect (the association between Cue B and the effect) to its indirect activation of the effect (the association from Cue B to Cue A multiplied by the association from Cue A to the effect).

As is typical for an associative model, the salience of a cue depends on whether it is present or absent. We assume that the salience of cue i depends on whether the cue is present ($s_i = \alpha$) or absent ($s_i = 0.0$), and the salience of cue j (which could be the effect) also depends on whether the cue is present ($s_j = \alpha$) or absent ($s_j = k_1$). The strength of the competition from the indirect activation of the effect is controlled by the parameter k_2. There is a final parameter k_3 that influences the extent to which a cue that has never been paired with the effect competes with other cues. In all of our simulations of the comparator hypothesis, we assume that context is ignored.

(*Appendices continue*)

Figure A1 shows the predicted asymptotic associations in recovery from overshadowing (AB+ A–) and backward blocking (AB+ A+) according to the comparator hypothesis. After the AB+ trials, the comparator hypothesis predicts that there will be limited responding to Cue B alone, because Cue B will activate the effect both directly and indirectly. On subsequent A– trials in recovery from overshadowing, responding to Cue B will approach 1.0 as the diminishing Cue A–effect association reduces the indirect activation of the effect. In contrast, given that the Cue A–effect association should already be near ceiling following the AB+ trials, subsequent A+ trials in backward blocking would have a limited influence on responding to Cue B. The comparator hypothesis predicts that responding to Cue B in backward blocking will continue to be limited but nonzero: Cue B will continue to activate the effect both directly and indirectly. Thus, like the Bayesian model, the comparator hypothesis predicts that people will be more certain about the causal influence of Cue B following recovery from overshadowing than following backward blocking.

Modified SOP Model

The modified SOP model (Dickinson & Burke, 1996) extends the SOP model (Wagner, 1981) to account for retrospective revaluations. The model represents each cue by a collection of elements. Because the model does not distinguish between cues and effects, the effect is also represented by a collection of elements. At any given time, an individual element will be in one of three states: the observed state, the expected state, or the inactive state. When a cue has not been observed recently and is not expected on the basis of a within-compound association, all of its elements will be in the inactive state. However, when a cue has been observed or is expected, some of its elements will be in these other states. Observing a cue causes some of its elements to move into the observed state, and the expectation of a cue (which is established through within-compound associations) causes some of its elements to move into the expected state. If a cue is both observed and expected, then we might find 40% of its elements in the observed state, 40% in the expected state, and 20% in the inactive state. Elements in the observed state eventually decay into the expected state and elements in expected state eventually decay into the inactive state. Thus, elements of an observed cue decay to the "expected" state even when the cue is not expected. For this reason, the observed and expected states are typically referred to more generally as the A1 and A2 states. However, the A1-to-A2 decay is often ignored when deriving the qualitative predictions of the model, so we adopt the more descriptive observed and expected terms.

In the modified SOP model, excitatory learning occurs between two cues to the extent that they are both in the observed state or both in the expected state. Inhibitory learning occurs between two cues to the extent that one is in the observed state, and the other is in the expected state. No learning occurs otherwise. For example, if a cue and the effect were presented together and no associations had been formed yet, some elements of the cue and some elements of the effect would both move into the observed state. This would lead to excitatory learning between the cue and the effect.

The upper half of Table A1 shows how the modified SOP explains recovery from overshadowing (AB+ A–) and backward blocking (AB+ A+). On AB+ trials, the modified SOP model learns that each cue is associated with the effect and that there is a within-compound association between Cues A and B. On subsequent A– or A+ trials, Cue B will be expected because of its within-compound association with Cue A. Hence, many elements of Cue B will be in the expected state. In recovery from overshadowing, the effect will be expected on the basis of its association with Cue A, so its elements will also enter the expected state. Since both the effect and Cue B will be in the expected state, the model predicts that people will become increasingly convinced that Cue B is a cause of the effect. However, for backward blocking, because the effect is observed *and* expected on the basis of its association with Cue A, its elements will enter both the observed and expected states. Because Cue B and the effect will be partly in the same state and partly in different states, this will induce conflicted (i.e., excitatory *and* inhibitory) learning. Thus, the modified SOP model predicts that Cue B should undergo strong learning on A– trials of recovery from overshadowing (i.e., people should become certain that Cue B causes the effect) and limited learning on the A+ trials of backward blocking (i.e., people should remain uncertain about the causal influence of Cue B).

Although the predictions of the modified SOP model are usually derived qualitatively, we consider a quantitative model to facilitate comparisons with the other models. When deriving the quantitative predictions of the model, we employ Aitken and Dickinson's (2005) implementation. Following Aitken and Dickinson (2005), we consider the parameters $P_{I \to A1}$ (the salience of an observed cue), l (the overall learning rate), ρ (the ratio of the observed- and expected-state learning rates), N_s (the number of elements per stimulus), and N_d (the rate of decay for elements in State A1).

(*Appendices continue*)

Experiments 1A and 1B

Figure 3 shows that the modified SOP model and the comparator hypothesis fail to explain the results in Experiments 1A and 1B. The predictions of the comparator hypothesis are discussed in the main text. We therefore focus on the predictions of the modified SOP model here.

Under the standard assumption that excitatory learning and inhibitory learning approximately counteract each other, the modified SOP model cannot explain the results of Experiment 1B. To see why, note that the modified SOP model predicts that learning regarding Cue C will be conflicted in recovery from preventive overshadowing but unambiguous in preventive backward blocking (see the bottom half of Table A1). This supports the erroneous prediction that the causal influence of Cue C will remain ambiguous in recovery from preventive overshadowing while potentially becoming unambiguous in preventive backward blocking.

Furthermore, the modified SOP model never—even with nonstandard assumptions—predicts the observed result that participants were more certain in both generative and preventive recovery from overshadowing than in generative and preventive backward blocking. As Table A1 makes clear, the modified SOP model predicts that recovery from overshadowing is more similar to preventive backward blocking than to preventive recovery from overshadowing. Depending on the assumptions that one makes about the relative strengths of excitatory and inhibitory learning, the modified SOP might explain either—*but not both*—the pattern of inference in Experiment 1A or the pattern of inference in Experiment 1B.

Experiment 2

As Figure 4 shows, both the comparator hypothesis and the modified SOP model predict substantial instability in the causal ratings for Cue A. According to the comparator hypothesis, the AB+ trials immediately establish Cue B as strong competitor for Cue A. The comparator hypothesis only predicts stable causal ratings in the second learning phase when the increase in the Cue A–effect association counteracts the increased competition from Cue B. Even when the parameters are set so, however, the B+ trials in Phase 3 will cause Cue B to become a still-stronger competitor, thereby predicting a later decrease in the causal ratings for Cue A.

The modified SOP predicts conflicted learning (i.e., both excitatory and inhibitory) whenever the effect is both expected and observed. In principle, this would allow the modified SOP to explain the stability of Cue A throughout the entire experiment and to account for the stability of Cue C during the third learning phase. In practice, however, excitatory and inhibitory learning will rarely offset each other perfectly. As Figure 4 shows, Aitken and Dickinson's (2005) implementation of the modified SOP predicts substantial changes in the causal ratings for Cues A and C.

Table A1
Selected Predictions of the Modified Sometimes-Opponent-Process Model for Generative and Preventive Variants of Recovery from Overshadowing and Backward Blocking

Condition	Activation states		Target-effect learning
	Target cue	Effect	
Generative variants			
RO (AB+ **A−**)	*E*	*E*	↑
BB (AB+ **A+**)	*E*	*O* + *E*	↑ ↓
Control (AB+)	*I*	*I*	none
Preventive variants			
pRO (A+ ABC− **AB+**)	*E*	*O* + *E*	↑ ↓
pBB (A+ ABC− **AB−**)	*E*	*E*	↑
Control (A+ AB−)	*I*	*I*	none

Note. RO = recovery from overshadowing; BB = backward blocking; pRO = recovery from (preventive) overshadowing; pBB = (preventive) backward blocking; *E* = expected state; *O* = observed state; *I* = inactive state; ↑ = excitatory learning, which in isolation would increase the associative strength; ↓ = inhibitory learning, which in isolation would decrease the associative strength. Learning is shown for trials that are displayed in boldface. The target cue is Cue B in the upper half of the table and Cue C in the lower half of the table.

(*Appendices continue*)

Appendix B

Model Fitting

The predictions of the associative models are parameter dependent. Whenever we derived the predictions of an associative model, we selected its parameters to provide the best fit to the experimental results. For the associative models other than the modified SOP model, we selected the parameter values by using the Nelder-Mead method, a gradient-descent procedure, to minimize the mean squared error of the predictions. To reduce the chances of identifying a local minimum, we performed this fitting procedure many times, with randomized initial parameters on each run. The fitting procedure was repeated at least 20 times for each fit, and the procedure consistently converged to one of a few local minima on each run.

Because the modified SOP model involves stochastic processes, its predictions vary slightly from run to run. This simulation noise limits the effectiveness of the gradient-descent procedures such as the Nelder-Mead method that propose small steps in the parameter space. We therefore fit the parameters of the modified SOP model with a different gradient-descent fitting procedure. For each iteration of the fitting procedure, we sequentially updated each parameter. To update a parameter, we varied it from 60% to 140% of its present value in 20% increments, while leaving the other parameters fixed, and then selected the value of the parameter for which the model provided the highest correlation with the experimental results (we maximize the correlation, rather than minimizing the mean squared error, because the associations in the modified SOP model have no natural maximum or minimum values). To estimate the model predictions more precisely, we averaged the predictions of 10 runs of the model at each parameter value. We iterated the fitting procedure until the model fit stabilized: We stopped updating the parameters when none of the parameter values changed during an iteration or when the model fit failed to improve on three consecutive iterations. Although the specific parameters found using this procedure differed when the fitting procedure was repeated multiple times, the final predictions were always qualitatively similar. The correlations between the final predictions and the causal ratings were nearly identical each time the fitting procedure was run: the standard deviation of the correlations across runs was less than .001 for Experiment 1 and was .011 in Experiment 2.

Note that although the R-W model has three parameters (α, β_1, and β_2), the model is fully specified by the values of $\alpha * \beta_1$ and β_2 / β_1: If the individual parameters are varied but these values remain constant, the model makes the same predictions. When reporting the parameter values of the R-W model, we therefore report the values of $\alpha * \beta_1$ and β_2 / β_1, rather than the values of the individual parameters. For similar reasons, for Van Hamme and Wasserman's (1994) learning rule and the noisy-logical associative model, we report the values of $\alpha_1 * \beta_1$, α_2 / α_1, and β_2 / β_1, rather than reporting the individual parameter values. Finally, we only report the value of k_3 for the comparator hypothesis when it influences the predictions of the model.

In Experiments 1A and 1B, the predictions for the associative models used the following best fitting parameters. R-W model: $\alpha * \beta_1 = 0.66$ and $\beta_2 / \beta_1 = 0.62$; Van Hamme and Wasserman's (1994) learning rule: $\alpha_1 * \beta_1 = 0.58$, $\alpha_2 / \alpha_1 = -0.37$, and $\beta_2 / \beta_1 = 0.67$; comparator hypothesis[4]: $k_1 = -0.75$, $k_2 = 0.49$, $k_3 \geq 40$, and $\alpha = 1.33$; the modified SOP model: $P_{I \rightarrow A1} = .37$, $l = 0.028$, $\rho = 0.004$, $N_s = 496$, and $N_d = 27$; the noisy-logical associative model: $\alpha_1 * \beta_1 = 0.32$, $\alpha_2 / \alpha_1 = -0.62$, and $\beta_2 / \beta_1 = 0.48$.

In Experiment 2, the predictions for the associative used the following best fitting parameters. The R-W model: $\alpha * \beta_1 = 0.33$ and $\beta_2 / \beta_1 = 0.76$; Van Hamme and Wasserman's (1994) learning rule: $\alpha_1 * \beta_1 = 0.35$, $\alpha_2 / \alpha_1 = -0.31$, and $\beta_2 / \beta_1 = 1.05$; the comparator hypothesis: $k_1 = -0.72$, $k_2 = 0.39$, and $\alpha = 0.51$; the modified SOP model: $P_{I \rightarrow A1} = .64$, $l = 0.07$, $\rho = 0.13$, $N_s = 1{,}187$, and $N_d = 80$; the noisy-logical associative model: $\alpha_1 * \beta_1 = 0.23$, $\alpha_2 / \alpha_1 = -0.39$, and $\beta_2 / \beta_1 = 1.41$.

[4] Predictions did not differ significantly for $k_3 \geq 40$. The values for k_3 and α lie outside what might be considered their natural range (between zero and one), but these values produced the best fit. When these values were constrained to be less than or equal to 1.0, the qualitative predictions of the model were similar, and the fit was only slightly worse ($MSE = 0.070$ compared to $MSE = 0.064$).

Received April 8, 2011
Revision received June 25, 2012
Accepted July 11, 2012 ■