# Uncertainty in causal inference: The case of retrospective revaluation

**Christopher D. Carroll (cdcarroll@ucla.edu)**
Department of Psychology, UCLA

**Patricia W. Cheng (cheng@lifesci.ucla.edu)**
Department of Psychology, UCLA

**Hongjing Lu (hongjing@ucla.edu)**
Department of Psychology, UCLA

**Abstract**

Since causal evidence is often ambiguous, models of causal learning should be able to represent uncertainty over causal hypotheses. Uncertainty is especially important in *retrospective revaluation* (the re-evaluation of ambiguous evidence in light of subsequent learning). We examine how a Bayesian model and an associative model (the modified SOP model of Dickinson & Burke, 1996) deal with this uncertainty. We tested the predictions of the models in an experiment with retrospective revaluation of preventive causes. Results were consistent with the predictions of the Bayesian model, but inconsistent with the predictions of the modified SOP model.

## Introduction

Causal evidence is often ambiguous and causal inference uncertain. When examining an isolated case of food poisoning, it is difficult to identify the meal – never mind the food item - that caused the illness. Uncertainty is especially salient in *retrospective revaluation* (when established but ambiguous evidence is re-evaluated after subsequent learning). We examine how Bayesian and associative models of causal learning represent and deal with ambiguous evidence in retrospective revaluation. Although Bayesian models naturally represent the uncertainty of causal inference from ambiguous evidence, associative models do not.

Examples of retrospective revaluation include *reduced overshadowing* and *backward blocking*. In both of these phenomena, there is one effect whose presence we denote as + and absence we denote as - and two cues that we will call cue A and cue B. In both reduced overshadowing and backward blocking, the initial evidence shows that the effect occurs after the presentation of both cues (AB+). This evidence is ambiguous because it could be that cue A alone causes the effect, cue B alone causes the effect, or that both cues A and B independently cause the effect. Of course, it is also possible that cues A and B interact to cause the effect, but we will not consider this possibility further. We assume that, due to parsimony, this explanation is only considered when the others are ruled out.

In reduced overshadowing, participants later learn that the effect does not occur after cue A is presented on its own (i.e., A- trials follow the AB+ trials). This new evidence suggests that cue A does not cause the effect. By conditional contrast or the process of elimination, this implies that cue B caused the effect on the AB+ trials. In backward blocking, the new evidence shows that the effect occurs when cue A is presented alone (i.e., A+ trials follow the AB+ trials). Since the knowledge that cue A causes the effect explains the AB+ trials, this new evidence should make it less likely that cue B causes the effect. However, it is still possible that cue B also causes the effect. Intuitively then, reduced overshadowing – which implies that cue B must cause the effect – should offer stronger evidence for re-evaluation than backward blocking.

This intuition is reflected in studies that have compared reduced overshadowing and backward blocking to a control condition (just AB+ trials). These studies have shown that reduced overshadowing is stronger and more robust than backward blocking (Corlett et al.., 2004; Larkin, Aitken, & Dickinson, 1998; see also Beckers, De Houwer, Pineno, & Miller, 2005; Lovibond, Been, Mitchel, Bouton, & Frohardt, 2003; but see Wasserman & Berglan, 1998; Wasserman & Castro, 2005).

In this paper, we consider how different models of causal reasoning explain reduced overshadowing and backward blocking. Our goals are two-fold. Firstly, we seek to provide a principled explanation of reduced overshadowing and backward blocking by representing uncertainty. In service of this goal, we formalize our intuitions in a Bayesian model of causal inference.

Secondly, we consider how associative models deal with retrospective revaluation. We focus on the modified SOP model (Dickinson & Burke, 1996) because it explains the observed asymmetry between reduced overshadowing and backward blocking. However, we will argue that the modified SOP model predicts this asymmetry for arbitrary reasons. Therefore, we tested the modified SOP and Bayesian models in a situation where they make competing predictions: the preventive analogs of reduced overshadowing (A+, ABC-, AB+) and backward blocking (A+, ABC-, AB-).

## A Bayesian model of retrospective revaluation

Bayesian models have been applied to retrospective revaluation in order to explain trial-order effects (e.g., Daw, Courville, & Dayan, 2008; Kruschke, 2008; Lu, Rojas, Beckers, & Yuille, 2008) and the influence of prior knowledge (e.g., Sobel, Tenenbaum, & Gopnik, 2004). These models, however, have not been contrasted with

associative models that were designed to explain retrospective revaluation. For this comparison, we adapt Griffiths & Tenenbaum's (2005) model of causal inference.

The model represents each possible causal explanation as a *causal graph* (e.g., Figure 1). In the causal graphs that we consider, a causal link can be generative, preventive, or non-existent. Since we assume that there are multiple cues and a single effect, a causal graph can be represented as a vector of causal links $\vec{l}$, 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. We provide each causal link with a weight that represents the strength of the causal relationship, and we represent these weights as a vector $\vec{w}$ where $0 \leq w_i \leq 1$ for each $w_i$.

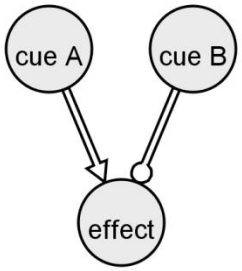


Figure 1: A causal graph where cue A causes the effect (as indicated by an arrow) and cue B prevents the effect (as indicated by a modified arrow terminating in a circle)

To represent a trial, we let the vector $\vec{c}$ denote the presence ($c_i = 1$) or absence ($c_i = 0$) of the cues and let $e$ denote the presence ($e+$) or absence ($e-$) of the effect.

To specify the probability of the effect, we need to define a generating function that describes how causes combine to produce the effect. We adopt the noisy-or and noisy-and-not generating functions, which can be derived from the assumptions of causal power (Cheng, 1997) for generative and preventive causation, respectively. Given the vectors $\vec{c}$, $\vec{l}$, and $\vec{w}$, 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 function, the probability of the effect is:

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

Then, given data $D$ that provides a frequency count $N(e, \vec{c})$ for each combination of the presence/absence of the effect and the cues, the probability of the data as a function of the causal graph and its weights is:

$$P(D \mid \vec{w}, \vec{l}) = \prod_{(e,\vec{c})} P(e \mid \vec{c}, \vec{l}, \vec{w})^{N(e,\vec{c})} \quad (2)$$

We assume a uniform prior distribution on $\vec{w}$ and define a prior distribution on $\vec{l}$ as shown in equation 3. For each causal link, we make the link generative with probability α, preventive with probability β, and nonexistent with probability 1 – α – β. We use α and β as model parameters. For a causal graph with $k$ generative causes, $j$ preventive causes, and $n$ cues, the priors are:

$$P(\vec{w}, \vec{l}) = P(\vec{w} \mid \vec{l}) P(\vec{l})$$
$$P(\vec{l}) = \alpha^k \beta^j (1 - \alpha - \beta)^{(n-k-j)} \quad (3)$$
$$P(\vec{w} \mid \vec{l}) \sim unif$$

From Bayes' theorem and our assumptions about the priors, we have

$$P(\vec{w}, \vec{l} \mid D) = \frac{1}{Z} P(D \mid \vec{w}, \vec{l}) P(\vec{w} \mid \vec{l}) P(\vec{l}) \quad (4)$$

The variable $Z$ represents a normalizing constant. The model can be used to answer questions about the strength of a causal link or about its existence and direction. To find the posterior probability of a set of causal weights (i.e., causal strengths), we can integrate equation 4 over the other causal weights and sum over the causal graphs.

The experiment in this paper, however, asks about the existence and direction of a causal link – not its strength. Therefore, we are more interested in the probability that a causal graph generated the data. This can be found by integrating over the causal weights.

$$P(\vec{l} \mid D) = \int P(\vec{w}, \vec{l} \mid D) d\vec{w} \quad (5)$$

To calculate the probability that a cue is causal, preventive, or noncausal, we sum the probabilities of each causal graph that contains the desired relationship. If we let $L$ be the set of causal graphs such that $l_i = x$ (where $x \in \{-1, 0, 1\}$ represents the existence and direction of the causal relationship), then:

$$P(l_i = x \mid D) = \sum_{\vec{l} \in L} P(\vec{l} \mid D) \quad (6)$$

Finally, to model causal judgments, we take the logit of this probability to obtain a measure of causal support, which is often viewed as a psychologically realistic measure of causal judgment (Griffiths & Tenenbaum, 2005):

$$\text{causal support} = \log\left(\frac{P(l_i = x \mid D)}{1 - P(l_i = x \mid D)}\right) \quad (7)$$

## Retrospective revaluation

To explain reduced overshadowing and backward blocking, we consider the causal graphs with two cues and one effect. Since we only allow causal relationships between cue A and the effect and cue B and the effect, this gives us 9 (i.e., $3^2$) causal graphs. We set the parameters such that the priors across the graphs are uniform (i.e., α = β = 1/3). When the model is given data where there are 4 trials of each type (e.g., 4xAB+ 4xA+ in the backward blocking condition), it can be used to generate a support measure for the hypothesis that cue B causes the effect. The model predicts that the difference between reduced overshadowing and a control (AB+) is larger than the difference between backward blocking and the control (see Table 1).

To understand these predictions, it is useful to consider the posterior distribution of the weights. First, we consider the joint posterior of cues A and B after the AB+ trials conditional on both links being generative (see Figure 2). This posterior suggests that there is considerable uncertainty over the weights of cues A and B. However, it also suggests a dependency between the weights of the cues: at least one

of the cues must be causal. If $w_a$ is small, then $w_b$ must be large. However, if $w_a$ is large, then there is still uncertainty over $w_b$. This dependency explains reduced overshadowing and backward blocking. If subsequent evidence indicates that cue A does not cause the effect (as is the case for reduced overshadowing), then cue B must. However, if subsequent evidence indicates that cue A causes the effect (as is the case for backward blocking), then the influence of cue B cannot be conclusively known.

Table 1: The causal support measure for the causal link between cue B and the effect for reduced overshadowing, control, and backward blocking

| Condition | support |
|---|---|
| reduced overshadowing (AB+, A-) | 5.02 |
| backward blocking (AB+, A+) | 0.09 |
| control (AB+) | 1.05 |

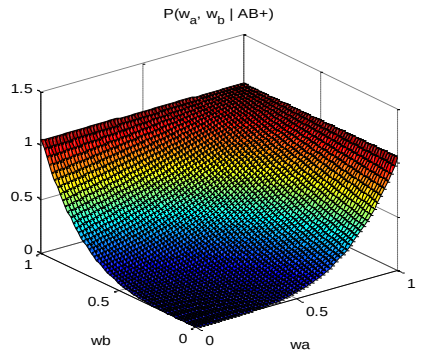


Figure 2: The joint posterior distribution of $w_a$ and $w_b$.

These predictions are reflected in the posterior weights of cue B alone in the different retrospective revaluation conditions (see Figure 3). In the reduced overshadowing condition, it is clear that cue B must cause the effect: there is almost no possibility that the weight from cue B to the effect is zero. On the other hand, there is considerable uncertainty about the weight of cue B in both the blocking and control conditions: neither excludes the possibilities that B is noncausal or that B is causal.

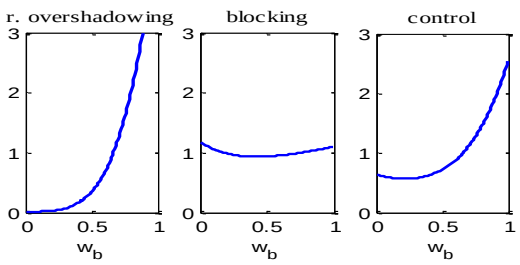


Figure 3: The posterior of the weights of cue B when cue B is a generative cause.

## Associative models

Retrospective revaluation is notoriously problematic for associative models, but two associative models have been developed to explain it: Van Hamme & Wasserman's (1994) modified RW (Rescorla-Wagner) model and the Dickinson & Burke's (1996) modified SOP model. The problem for standard associative models is that they only learn about present cues. This precludes an explanation of reduced overshadowing and backward blocking, where learning about cue A leads participants to revise their beliefs about the absent cue B. To surmount this difficulty, the modified RW model and the modified SOP model utilize *within-compound associations:* associations formed between simultaneously-presented cues. On the initial AB+ trials, these models learn an association between cues A and B. Later, the within-compound associations are used to recall associated cues that are absent on the trial, allowing the model to learn about them. If an A+ trial followed, the models would identify the absent cue B as an expected cue and would use this identification to support re-evaluation.

Although within-compound associations allow the models to learn about absent cues, it is not clear whether they offer a genuine representation of uncertainty.

Since the modified RW model incorrectly predicts that backward blocking will be at least as strong as reduced overshadowing (see Larkin et al., 1998 for a detailed explanation), we focus on the modified SOP model.

### The modified SOP model

In the modified SOP model, there are three activation states: the A1 (observed), A2 (expected), and I (inactive) states. Each cue is represented by a node that is made up of many elements, so a node can be in more than one activation state. For example, if a cue were presented on a trial and it was expected on the basis of within-compound associations, there might be 40% of its elements in the A1 state, 40% in the A2 state, and 20% inactive. Excitatory learning occurs between two nodes to the extent that they are both in the A1 state or both in the A2 state. Inhibitory learning occurs between two nodes to the extent that one is in the A1 state and the other is in the A2 state. No learning occurs otherwise.

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. When cue A is presented alone, the within-compound association between cues A and B leads cue B to enter the A2 activation state (see Table 2). The state of the effect depends on the type of retrospective revaluation. For reduced overshadowing, the effect is expected but absent, so it will enter state A2. This puts the effect in the same state as cue B, so learning will be exclusively excitatory. For backward blocking, the effect is both expected and present, so it will enter states A1 and A2. Since this means that cue B and the effect will be partly in the same state and partly in a different state, there will be conflicted learning that is both excitatory and inhibitory. Therefore, the modified SOP model predicts the oft-

observed asymmetry between reduced overshadowing and backward blocking. When compared to a control condition, the modified SOP model predicts that reduced overshadowing will be a stronger and more robust effect than backward blocking.

Table 2: Activation states and learning during retrospective revaluation. The ↑ symbol indicates excitatory learning (an increase in associative strength) and the ↓ symbol indicates inhibitory learning (a decrease in associative strength).

| condition | Cue B | Effect | B-effect learning |
|---|---|---|---|
| r. overshadowing (A-) | A2 | A2 | ↑ |
| b. blocking (A+) | A2 | A1 and A2 | ↑↓ |
| control | - | - | none |

However, these predictions seem arbitrary: the modified SOP model predicts both excitatory and inhibitory learning whenever the effect is both present and expected, but it is not clear why this should be the case. To test the modified SOP model, we designed an experiment where its predictions diverged from those of the Bayesian model.

## Method

To test the predictions of the modified SOP model, we examined the preventive analogs of reduced overshadowing (i.e., A+, ABC-, AB+) and backward blocking (i.e., A+, ABC-, AB-). Until the final AB+ or AB- trials, the evidence suggests that cue A causes the effect and that either cue B alone prevents the effect, cue C alone prevents the effect, or that cues B and C prevent the effect. Like its generative analog, preventive reduced overshadowing eliminates two of these explanations by showing that cue B does not prevent the effect. By the process of elimination, one would infer that cue C must have been responsible for preventing the effect on the ABC- trials. The AB- trials in backward blocking show that cue B prevents the effect, but these trials do not fully clarify the influence of cue C: it is still possible that C prevents the effect, and it is still possible that it does not. Preventive reduced overshadowing should be a stronger and more robust effect than preventive backward blocking.

The modified SOP model predicts the opposite. It predicts that learning is conflicted whenever the effect is both present and expected (as it is during reduced overshadowing AB+ trials), but that learning is clear whenever the effect is expected but absent (as it is during backward blocking AB- trials). According to the modified SOP model, preventive reduced overshadowing should be weaker and less robust than preventive backward blocking (see Table 3).

For our experimental task, we used a cover story where participants were asked to discover which foods cause and prevent allergic reactions in medical patients. We manipulated the retrospective revaluation condition (preventive reduced overshadowing, preventive backward blocking, reduced overshadowing control, and blocking control) within-subjects. We also manipulated expectations about the probability that a randomly selected fruit would prevent an allergic reaction. Bayesian models have a mechanism for integrating prior knowledge and evidence from observations, and we manipulated expectations to assess whether prior knowledge influenced the participants.

Table 3: Predicted changes in associative strength according to the modified SOP model for the preventive analogs of reduced overshadowing and backward blocking.

| Preventive analog | Cue C | Effect | C-effect learning |
|---|---|---|---|
| r. overshadowing (AB+) | A2 | A1 and A2 | ↑↓ |
| b. blocking (AB-) | A2 | A2 | ↑ |
| control | - | - | none |

### Participants

Twenty-four undergraduates at the University of California, Los Angeles participated for course credit. The participants were randomly assigned to a infrequent ($n$ = 7), occasional ($n$ = 9), or frequent ($n$ = 8) prevention condition.

### Materials

We selected icons that pictorially represented 21 different fruits.

### Procedure

At the beginning of the experiment, participants were asked to take the perspective of allergists specializing in patients who have fruit allergies. They were informed that fruit allergies can be both caused and prevented in these patients. That is, some fruits might cause an allergic reaction in a patient, but other fruits might prevent an allergic reaction.

Participants were told that they would read through the "fruit journals" of patients. They were informed that a fruit journal lists the fruits that a patient ate on a given day, and also records whether the patient had an allergic reaction.

Each experimental trial corresponded to the record for one day in the fruit journal. A trial began by displaying the icons and names of whichever fruits the patient ate on that day. These icons were displayed alone for 1.5 seconds, 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" meant that the patient did not have a reaction and a frowning face with the text "allergic reaction" meant that the patient had a reaction. The fruits and cartoon face were displayed together for 2.0 seconds before the trial ended.

Participants read the fruit journals of five different patients. The journal of the first patient was used to manipulate the priors. The other four journals represented the four retrospective revaluation conditions. The fruits

were randomly mapped to the different fruit journals, and each fruit appeared in exactly one fruit journal.

When the first patient was introduced, participants were told the approximate probability that a fruit prevents allergic reactions (the bracketed phrases were selected according to the infrequent, occasional, or frequent priors conditions):

> *As is often the case with fruit allergies, a small number of fruits caused the patient's allergic reaction, [very few / some / many] prevented it, and [many / some / very few] did nothing.*

The first fruit journal provided evidence for this claim. The patient experienced an allergic reaction after consuming one of the fruits alone, but the other four fruits in the journal did not cause the patient to experience an allergic reaction. Zero, two, or four of the other fruits prevented the allergic reaction (in the infrequent, occasional, and frequent priors conditions, respectively). This was demonstrated by showing, for each of the other fruits, whether the patient had an allergic reaction after consuming that fruit and the causal fruit at the same time.

To familiarize the participants with the causal questions, the participants were then asked whether each fruit in the first journal caused, prevented, or did nothing to influence the patient's allergic reactions. Participants responded on a sliding scale running from -6 to 6 where -6 was labeled "definitely prevents", -3 was labeled "maybe prevents", 0 was labeled "neither", 3 was labeled "maybe causes", and 6 was labeled "definitely causes."

After answering questions about the influence of fruits on the first patient, participants viewed, in random order, a fruit journal for each retrospective revaluation condition. In each journal, the trials were divided into three stages, and the data for each stage are shown in Table 4. Each of the listed patterns was shown four times (e.g., fruit A caused an allergic reaction four times in stage 1). Within each stage, the trials were presented in a random order.

Table 4: The data (by retrospective revaluation condition)

| condition | stage 1 | stage 2 | stage 3 |
|---|---|---|---|
| reduced overshadowing | A+ B- C- D- | A+ ABC- | A+ AB+ |
| backward blocking | A+ B- C- D- | A+ ABC- | A+ AB- |
| control (rOS) | A+ B- C- D- | A+ ABC- | A+ AD+ |
| control (BB) | A+ B- C- D- | A+ ABC- | A+ AD- |

Following the presentation of the data for each retrospective revaluation condition, participants were asked to report whether each fruit caused, prevented, or did nothing to influence the patient's allergic reactions. The response scale was identical to the scale that was used in the first fruit journal.

## Results

The ratings for cue C are shown in Figure 4. The predicted asymmetry between reduced overshadowing and backward blocking was found. Compared to its control, reduced overshadowing had a substantial influence: it led participants to be much more certain that cue C prevented allergic reactions. Ratings for cue C did not differ substantially between the backward blocking and control conditions. The priors manipulation did not seem to substantially influence the causal ratings.

An ANOVA confirmed that the retrospective revaluation condition influenced causal ratings, $F(3, 63) = 23.84$, $p < .001$, and that there was no effect of the priors manipulation, $F(2, 21) = 0.29$, $p = .75$, or interaction between the priors condition and retrospective revaluation condition, $F(6, 63) = 0.57$, $p = .75$. Planned comparisons indicated that the effect of retrospective revaluation condition was driven by the difference between reduced overshadowing and its control, $t(23) = 6.30$, $p < .001$, and not by the difference between blocking and its control, $t(23) = 0.94$, $p = .36$.

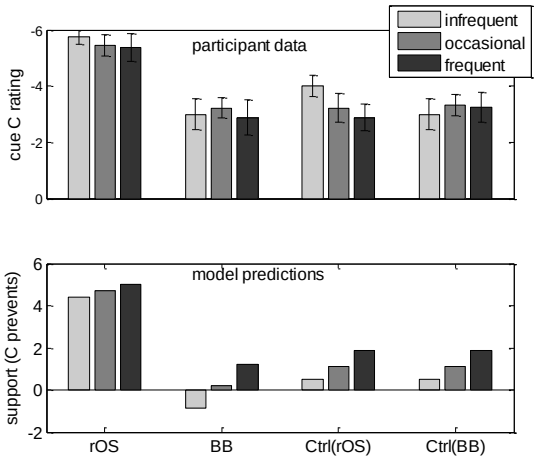


Figure 4: Causal ratings and Bayesian model predictions for cue C by retrospective revaluation condition and the prior likelihood of prevention. On both graphs, higher points on the y-axis correspond to greater certainty that cue C prevents allergic reactions. (rOS = reduced overshadowing, BB = backward blocking, Ctrl = control)

To derive the predictions of the model, we set $\alpha = .2$ and then set $\beta$ depending on the priors condition ($\beta = .2$ for infrequent, $\beta = .4$ for occasional, and $\beta = .6$ for frequent). The predictions of the model are shown for each condition in Figure 4. The model offered a good quantitative fit to the data, $r = -.87$.

## Discussion

The results clearly contradict the predictions of the modified SOP model. Preventive reduced overshadowing was a much stronger effect than preventive backward blocking. The Bayesian model predicts this finding, and also offers a principled justification for its prediction.

The priors manipulation did not influence the participants' causal ratings, but the interpretation of this finding is unclear. The Bayesian model predicts a limited effect of the priors manipulation, and the small number of participants per condition limited the experiment's statistical power. Furthermore, since the prior frequencies were merely manipulated verbally, the manipulation may have been too weak. Other research has shown that priors can influence causal judgment (e.g., Sobel, Tenenbaum, & Gopnik, 2004).

A final possibility is that the participants only represented approximate probabilities. Participants may have categorized the probability of causation by tracking whether a causal link *definitely, maybe,* or *definitely does not* exist. Consistent with this possibility, participants did not seem to differentiate between different degrees of *maybe* (e.g., see Figure 4).[1]

The modified SOP model predicts the relative size of reduced overshadowing and backward blocking, but the preventive analogs of these findings illustrate that it does so for the wrong reasons. In both the modified RW model and modified SOP models, within-compound associations make a poor substitute for a genuine representation of uncertainty. Other associative models that use within-compound associations may be capable of explaining these results (e.g., Denniston, Savastano, & Miller, 2001), so further experimentation is necessary. However, the results of this experiment raise serious questions about whether within-compound associations offer a genuine representation of uncertainty. As instantiated by the modified SOP model, they clearly do not.

## Acknowledgments

The preparation of this article was supported by AFOSR FA 9550-08-1-0489.

## References

Beckers, T., De Houwer, J., Pineno, O., & Miller, R. (2005). Outcome additivity and outcome maximality influence cue competition in human causal learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 31 (2),* 238-249.

Corlett, P. R., Aitken, M. R. F., Dickinson, A., Shanks, D. R., Honey, G. D., Honey, R. A. E., Robbins, T. W., Bullmore, E. T., & 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.

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.

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

Dickinson, A. & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgements. *The Quarterly Journal of Experimental Psychology, 1996, 49B (1),* 60-80.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51,* 334-384.

Kruschke, J. K. (2006). Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychological Review, 113 (4),* 677-699.

Larkin, M. J. W., Aitken, M. R. F., & Dickinson, A. (1998). Retrospective revaluation of causal judgments under positive and negative contingencies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24 (6),* 1331-1352.

Lovibond, P. F., Been, S., 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 (1),* 133-142.

Lu, H., Rojas, R. R., Beckers, T., & Yuille, A. (2008). Sequential causal learning in humans and rats. *Proceedings of the Twenty-ninth Annual Conference of the Cognitive Science Society*

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.

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

Wasserman, E. A., & Castro, L. (2005). Surprise and change: Variations in the strength of present and absent cues in causal learning. *Learning & Behavior, 33 (2),* 131-146.

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

[1] Interestingly, causal support does something similar. Rather than using the untransformed probability of a causal link (equation 6), causal support transforms this probability with the logit function. The logit function deemphasizes differences in moderate probabilities (i.e., those near .5) and emphasizes differences in extreme probabilities.

It is worth noting, however, that the logit transformation would not save the modified RW and the modified SOP models. Even when augmented with a logit transformation, these models fail to explain the results. The modified RW model learns that cue C is non-causal in preventive backward blocking more quickly than it learns that cue C is preventive in preventive reduced overshadowing. If anything, the logit transformation would highlight this failing. The modified SOP model makes predictions in the wrong direction. As a monotonic transformation, the logit function preserves the direction of these incorrect predictions.