Uncertainty and dependency in causal inference

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

When inferring causal relationships, people are often faced with ambiguous evidence. Models of causal inference have taken different approaches to explain reasoning about such evidence. One approach – epitomized by Bayesian models of causal inference – defers judgment by representing uncertainty across multiple explanations. Another approach – usually adopted by associative models – approximates uncertainty by positing within-compound associations, a special type of association that forms between simultaneously presented cues. Although these approaches explain many of the same experimental findings, we note some limitations of the latter approach. Within-compound associations form whenever two cues are presented simultaneously – even when the causal influences of the cues are already known. Since associative models use within-compound associations to modify beliefs about one potential cause when learning about another, associative models therefore predict that cues with known causal influences can have their influence revised as a result of being presented with other cues. In two experiments, we tested the predictions of the two approaches. The results were consistent with models that represent uncertainty across multiple explanations and inconsistent with models that use within-compound associations.

Keywords: causal reasoning; causal inference; uncertainty; associative models; Bayesian models

Introduction

Everyday causal inference often requires reasoning about ambiguous evidence. Consider a reasoner who is trying to explain events such as a recent illness, a lapse in a friendship, or a car accident. Each of these events has many possible explanations, and, in many cases, the reasoner will not be able to identify the correct explanation with certainty.

While there are different ways to respond to ambiguous evidence, the most reasonable response involves deferring judgment by representing the uncertainty associated with the evidence. That is, rather than committing to a single explanation prematurely, a reasoner presented with ambiguous evidence should distribute his or her belief across the possible explanations in accordance with the plausibility of each explanation. It has commonly been observed that while Bayesian models exemplify this sort of approach, associative models adopt another approach (e.g., Courville, Daw, Gordon, & Touretzky, 2003; Kruschke, 2008; Lu, Rojas, Beckers, & Yuille, 2008; Sobel, Tenenbaum, & Gopnik, 2004).

Most of the research that investigates reasoning about ambiguous evidence focuses on situations where there is an *inferential dependency* such that learning about whether one cue causes the effect provides information about whether other cues cause the effect. Consider, for example, a situation where the effect occurs in the presence of two possible causes (we write this as AB+, letting letters represent the potential causes and +/– represent the presence/absence of the effect). This ambiguous evidence establishes an inferential dependency between cues A and B because subsequent learning about cue A can provide information about cue B and vice versa (e.g., if shown A– trials, a reasoner would probably conclude that cue B definitely causes the effect). As we will see later, Bayesian models can explain this inferential dependency – as well as other inferential dependencies – by distributing belief across multiple explanations.

The evidence for the superiority of this approach, however, is less strong than one might expect: some associative models – which do not seem to defer judgment or represent uncertainty – explain the same experimental findings. At the very least, associative models are able to approximate a genuine representation of uncertainty and handle some types of ambiguous evidence. In the present paper, we present (1) an analysis of the weakness of the associative approximation of uncertainty and (2) an empirical test based on the analysis that clearly differentiates the two approaches.

Associative models

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

$$\Delta V_i = s_i s_e (T - \sum_i V_i) \quad (1)$$

In this learning rule, s_i represents the salience or learning rate for cue i when it is present ($s_i = \alpha$) or absent ($s_i = 0$), s_e represents the salience of the effect ($s_e = \beta$), T represents the presence ($T = 1$) or absence ($T = 0$) of the effect, and V_i represents the prior association between cue i and the effect. The summation, which occurs over all cues present on a given trial, represents the predicted strength of the effect. The difference between T and the summation therefore

represents the prediction error (observed – expected), and the model modifies the association between the cue and the effect in order to reduce this error on future trials.

The RW model accounts for notable experimental findings such as forward blocking (A+ AB+). Compared to a control condition without the initial A+ trials (i.e., AB+ alone), blocking produces a weaker association between cue B and the effect. The RW model explains this finding because it learns a strong association between A and the effect during the A+ trials. Consequently, the prediction error on the AB+ trials will be small, leaving little room for learning an association between cue B and the effect.

When presented with ambiguous evidence, however, the RW model fails to predict the existence of inferential dependencies. Consider the model predictions for backward blocking (AB+ A+) and recovery from overshadowing (AB+ A–), the most commonly demonstrated dependencies. Since the RW 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.

Associative models have been proposed that can explain these findings, including the Van Hamme and Wasserman (1994) model, the comparator hypothesis (Denniston, Savastano, & Miller, 2001; Miller & Matzel, 1988; Stout & Miller, 2007), and the modified SOP model (Dickinson & Burke, 1996). These models explain inferential dependencies by positing *within-compound associations*, associations formed between cues that are presented on the same trial. These associations are used to support learning about absent cues. Since problems with the modified SOP model have been considered previously (Carroll, Cheng, & Lu, 2010), we focus on the Van Hamme and Wasserman model and the comparator hypothesis in this paper.

The Van Hamme and Wasserman model

The Van Hamme and Wasserman (1994) model modifies the RW model by (1) imposing within-compound associations between the cues and (2) positing a negative learning rate for expected but absent cues. In short, while the Van Hamme and Wasserman model also uses Equation 1 to update the associations, it assigns different saliences to the cue depending on whether it is present ($s_i = \alpha_1$), absent but expected ($s_i = \alpha_2$ where α_2 is negative), or absent and unexpected ($s_i = 0.0$). The salience of the effect also varies as a function of its presence ($s_e = \beta_1$) and absence ($s_e = \beta_2$).

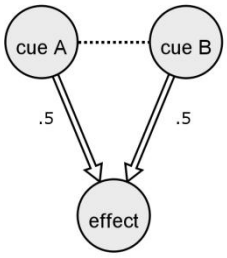

Figure 1: The asymptotic associations of the Van Hamme and Wasserman model on AB+ trials. The dashed line represents a within-compound association.

Figure 1 shows the asymptotic associations of the Van Hamme and Wasserman model during the AB+ phase of backward blocking or recovery from overshadowing. The modifications create a dependency between cues A and B. If shown an A+ trial after the AB+ trials, for example, the A-effect association will increase and the B-effect association will decrease.

To derive the quantitative predictions of the Van Hamme and Wasserman model, we follow Wasserman and Castro (2005) by letting $\alpha_1 = .7$, $\alpha_2 = -.4$, $\beta_1 = .5$ and $\beta_2 = .4$.

The comparator hypothesis

According to the comparator hypothesis (Denniston, Savastano, & Miller, 2001; Miller & Matzel, 1988; Stout & Miller, 2007), the direct activation of the effect from a cue (its association) is compared to the indirect activation of the effect from the cue (which is approximately equal to the product of the associations along an indirect path to the effect).[1] Figure 2 shows the asymptotic associations of the comparator hypothesis for AB+ trials. Since the direct and indirect activations of the effect by cue B are both large following the AB+ trials, the comparator hypothesis predicts that responding to cue B alone should be limited after these trials. Subsequent learning could influence this prediction, however. On A– trials, for example, the A-effect association would decrease, attenuating cue B's indirect activation of the effect. The comparator hypothesis predicts that cue B will be viewed as a stronger predictor of the effect after the A– trials, thereby explaining recovery from overshadowing.

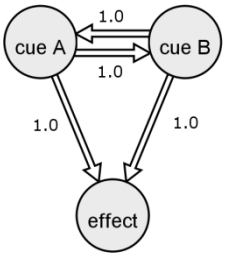

Figure 2: The asymptotic associations of the comparator hypothesis following AB+ trials.

The comparator hypothesis updates its associations using an equation more similar to Hull's (1952) than to the RW and Van Hamme and Wasserman equations:

$$\Delta V_{i,j} = s_i s_j (T - V_{i,j}) \quad (2)$$

Because Equation 2 calculates the error relative to the prediction of a single association ($V_{i,j}$) rather than relative to a sum of associations, the AB+ trials of backward blocking and recovery from overshadowing lead to associations that

[1] When there are two or more comparison cues, the comparison mechanism is more complicated. Because none of the present experiments involve higher-order comparisons, we can ignore these complications.

asymptotically approach 1.0 (see Figure 2). 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 ($s_j = k1$).

A final[2] parameter *k2* controls the strength of the competition from the indirect activation of the effect. Following Stout and Miller (2007), we set $\alpha = .7071$, $k1 = .1768$, $k2 = .9$ as the default parameters of the model. In all of our simulations, we assume that context is ignored.

Bayesian models of causal inference

Many Bayesian models of causal inference have been proposed. These models often represent possible explanations of the data as causal graphs. In these models, inferential dependencies arise from the ability of the model to distribute belief across multiple explanations. Consider Figure 3, which shows what a Bayesian model that assumes deterministic causation might infer from AB+ trials. The AB+ trials will concentrate belief on the explanations where at least one of the cues causes the effect. This belief distribution sets up an inferential dependency. To see why, observe that the probability that cue B causes the effect given that cue A causes the effect is 0.33/(0.33+0.33) = 0.5, whereas the probability that cue B causes the effect given that cue A does not cause the effect is 0.33/(0.33+.00) = 1.0. This suggests a dependency where learning whether or not cue A causes the effect will influence the model's beliefs about cue B.

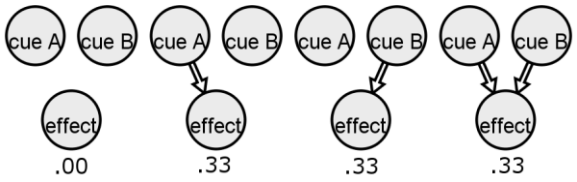

Figure 3: The predictions of a simple Bayesian model when given AB+ evidence. The numbers are the posterior probabilities of the explanations given the data, P(G|D).

We adopt a Bayesian model of causal inference that extends Griffiths and Tenenbaum's (2005) model. Provided with a set of cues, the model considers all of the causal graphs where each cue is either a cause of the effect or does not influence the effect. Given some data *D*, the model uses Bayes theorem to calculate the posterior probability of each graph G:

$$P(G|D) = \frac{P(D|G)P(G)}{P(D)} \quad (3)$$

The likelihood function, P(D|G), can be specified by assuming that causes produce their effects in accordance with causal power theory (Cheng, 1997). We assume uniform priors across both the causal graphs and the causal weights that the likelihood function uses to represent the strengths of the causal relationships. For a more detailed description of the model, see Carroll, Cheng, and Lu (2010).

Distinguishing the models

Due to the assumption that within-compound associations form whenever two cues are paired together, the associative models predict inferential dependencies in some counterintuitive situations. Consider what these models will predict when A+ trials are followed by AB+ trials. The models will establish within-compound associations during the AB+ trials, suggesting that subsequent learning about cue B might lead people to make an inference about cue A. According to the Bayesian model, on the other hand, the unambiguous A+ trials will concentrate all or almost all of the posterior probability on explanations where cue A is a cause of the effect. Subsequent evidence about other cues may change how belief is distributed between these explanations, but it is not likely to shift belief to other explanations. Two experiments tested whether, as the associative models predict, the simultaneous presentation of two cues might lead people make inferences about cues for which causal influence was unambiguously established on previous trials.

Experiment 1

The comparator hypothesis predicts that cues will become competitive as soon as there is a within-compound association between them. Consider a situation where a reasoner learns that two cues predict the effect separately (A+ B+) before learning that the cues predict the effect as a compound (AB+). The comparator hypothesis predicts that the initial A+ and B+ trials will establish strong cue-effect associations and that the subsequent AB+ trials will put the cues in competition by establishing a within-compound association between them. Therefore, participants should become less certain that either cue A or cue B causes the effect after observing AB+ trials.

Experiment 1 contrasted these predictions with the predictions of the Bayesian model through the experiment design shown in Table 1. The two causes and one cause conditions present situations that may be problematic for the comparator hypothesis. The causal influences of the cues in these conditions are unambiguous after phase 1, but within-compound associations will form during phase 2. These within-compound associations have the potential to decrease responding to the cues. Two control conditions were also included in the experiment. The competition control was included to confirm that the experiment was adequate to establish competition between cues in at least some circumstances. Without this control, the comparator hypothesis could account for a lack of competition in the experimental conditions by setting the parameter that controls the amount of competition to zero. A second control was included in order to control for forgetting.

[2] The model's fourth parameter *k3* does not influence on the predictions of the model in any of the present experiments.

Table 1: The data presented in Experiment 1.

condition	phase 1	phase 2
two causes	A+ B+	AB+
one cause	C– D+	CD+
competition control		EF+
forgetting control	G+ H–	

Method

Participants

Eighteen undergraduates at the University of California, Los Angeles participated for course credit.

Materials and Procedure

The experimental instructions informed the participants that they would be attempting to diagnose the fruit allergies of a patient at the hospital. Participants were told that the diagnoses would be made by reviewing the patient's "fruit journal." The fruit journal provided a daily log of the fruits that the patient ate and of his allergic reactions.

Table 1 summarizes the content of the fruit journal. Since we wanted to assess how the participant's causal beliefs changed over the course of the experiment, we presented the fruit journal in two separate learning phases. Participants reported their causal beliefs after each learning phase.

Within a learning phase, there were five trials for each trial-type (i.e., in the first phase, there were five A+ trials, five B+ trials, and so on), and the trials were presented in random order. Each trial began by displaying the icons and labels of whichever fruits the patient ate on that day. The icons and labels of the fruits were displayed alone for 1.5 seconds, at which point an 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 an allergic reaction and a frowning face with the text "allergic reaction" indicated that the patient had an allergic reaction. The trial concluded after the fruit or fruits and cartoon face were displayed together for 2.0 seconds.

After each learning phase, participants reported their causal beliefs by answering questions such as:

> *Suppose that on a given day, coconuts are the only fruit that the patient eats. Do you think that the patient will have an allergic reaction on that day?*

The participants responded on a slider with seven tick marks. The leftmost tick was labeled "definitely not," the middle mark was labeled "maybe," and the rightmost tick was labeled "definitely." Responses were coded as integers ranging from 1 ("definitely not") to 7 ("definitely").

Results and Discussion

Figure 4 shows the participant's causal ratings and the predictions of the Bayesian model and comparator hypothesis. The most informative comparisons are between the final ratings for the cues in the experimental conditions (i.e., cues A, B, C, and D) to the final ratings for the relevant forgetting control cue (i.e., cue G for the causal cues; cue H for the noncausal cues). Because the comparator hypothesis predicts that competition will develop between the experimental cues during phase 2, it predicts that the final causal ratings for the experimental cues will be lower. This was not the case. No significant differences were found between the final causal ratings for cue G and the final causal ratings for cue A, $t(17) = 0.79$, $p = .44$, cue B, $t(17) = 0.11$, $p = .91$, or cue C, $t(17) = 0.77$, $p = .45$. Similarly, the difference between the causal ratings for cue H and cue D was also non-significant, $t(17) = 1.49$, $p = .15$.[3] The failure to find differences between these cues is not due to a simple lack of statistical power: the participants clearly distinguished between different cues in phase 1, $F(5, 85) = 126.01$, $p < .001$, and phase 2, $F(7, 119) = 30.51$, $p < .001$. The Bayesian model correctly predicts the relative stability of the experimental cues.

As one might expect, therefore, the Bayesian model provided a better fit to the data ($r = .98$) than the comparator hypothesis ($r = .59$). To investigate whether the comparator hypothesis could explain the results when other parameter settings were adopted, we searched for the parameters that maximized the correlation between the model predictions and the causal ratings across both phases. With the best-fitting parameters, the model offered a much better fit ($r = 0.99$ with $\alpha = .30$, $k1 = .08$; $k2$ did not influence the model predictions). The better fit was achieved by slowing down the learning rate (α). With a slower learning rate, the associations between the cues and the effect did not approach asymptote on the trials in phase 1. Consequently, the associations continued to increase in phase 2. The best-fitting parameters adjusted the magnitude of this increase so that it exactly offset the increased competition that arises through the formation of the within-compound association.

Although post-hoc better-fitting parameters made the comparator hypothesis's predictions correlate well with the results, there are reasons to be suspicious of this adjustment. First, in the better fitting model, the predicted cue-effect associations in phase 1 (.37) are far from asymptote, making it awkward to explain why participants viewed the causal cues (cues A, B, D, and G) as "definite" causes of the effect. Furthermore, the model only predicts a stable cue A-effect association under very specific parameter settings. The causal influence of cue A will only be stable during the AB+ trials when the increase in the direct activation of the effect from cue A is exactly offset by the increase in the indirect activation via cue B. This delicate balance would be difficult to maintain across many situations.

[3] Rather than comparing the final ratings for the experimental and control cues, one could compare the initial and final ratings for the experimental cues. However, the small differences between these ratings were not statistically significant and could be attributed to forgetting in any case.

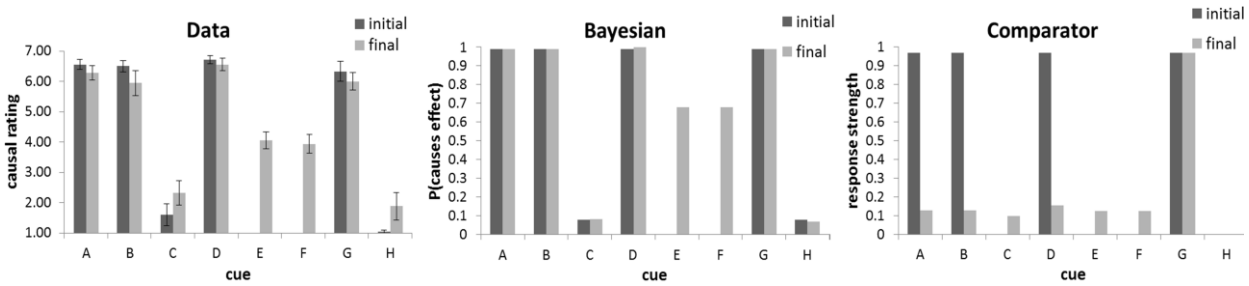

Figure 4: The data from Experiment 1 and the predictions of the models.

Experiment 2

Experiment 2 sought to find a clearer refutation of the comparator hypothesis and to modify the experimental procedure so that the Van Hamme and Wasserman model also predicts inferential dependencies. Because the Van Hamme and Wasserman model predicts that within-compound associations are only utilized when there is an expected but absent cue, testing the predictions of the model involves presenting one of the cues in isolation after a within-compound association has been formed.

Method

Participants

Eleven undergraduates 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 Experiment 1. The data were presented in three phases rather than two, and Table 2 shows the presented data. We also altered the experimental procedure in an attempt to limit the influence of forgetting across the phases. Rather than presenting data about the allergic reactions of a single patient to many fruits, we presented data about the allergic reactions of three different patients (one for each experimental condition). Participants viewed all of the data for one patient before moving on to the next patient. As was the case in the previous experiment, participants reported their causal beliefs after each phase.

Table 2: The presented data. The bold trials involved cues whose associations with the effect were analyzed. The other cues were only included as controls.

condition	phase 1	phase 2	phase 3
two causes	**A**+ G–	**AB**+	**B**+
one cause	**C**– H+	**CD**+	**D**+
recovery from overshadowing	I+ J–	**EF**+	**F**–

Results and Discussion

Figure 5 shows the causal ratings for the cues across the phases, as well as the predictions of the Bayesian model and the associative models with the best-fitting parameters. The model predictions differ most informatively for cues A, C, and E. Across the learning phases, the causal ratings for cues A and C were much more stable than the causal ratings

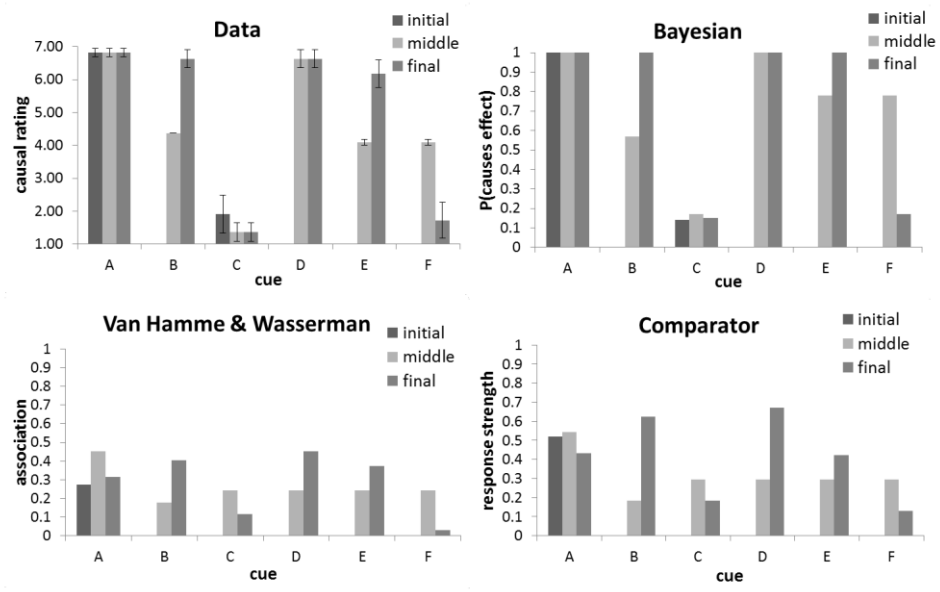

Figure 5: Results and predictions for Experiment 2.

for cue E. The Bayesian model predicts the relative stability of these ratings. The associative models do not, even when the parameters of the associative models were selected to maximize the correlation with the causal ratings. Consequently, the parameter-free Bayesian model (r = .98) provides a better fit to the data than the Van Hamme and Wasserman model (r = .81 with $\alpha_1\beta_1$ = .06, α_2 / α_1 = -.60, β_2 / β_1 = 5.53) and the comparator hypothesis (r = .81 with α = .37, $k1$ = .41, and $k2$ = .84).

Planned comparisons confirmed that the ratings for cues A and C were stable across phases 2 and 3 (in fact, none of the participants gave these cues different causal ratings in the two phases) and that there was a clear change in the ratings for cue E across these phases, $t(10) = 5.04$, $p < .001$. Since the Van Hamme and Wasserman model predicts (1) that a within-compound association will form on the AB+ trials and (2) that the B+ trial will be very surprising, it incorrectly predicts that beliefs about cue A will change dramatically during phase 2. The comparator hypothesis can only predict stable ratings for cue A on the AB+ trials if the learning rate is slow, but a slower learning rate insures that the B-effect association will still be increasing during the B+ trials. It is impossible for the comparator hypothesis to predict the stability of the causal ratings for cue A on both the AB+ and the B+ trials.

General Discussion

Associative models predict that inferential dependencies can arise whenever two cues are simultaneously presented. In situations where the causal influence of one of the cues is already known with near-certainty, this prediction can be distinguished from the predictions of Bayesian models, which will not predict inferential dependencies in such circumstances. The results in Experiments 1 and 2 favor the Bayesian model over the associative models.

These experiments suggest that any model of causal inference should represent uncertainty by distributing belief across multiple explanations. A model that does so – whether through probabilistic inference, propositional reasoning, or other mechanisms – will be able to explain the appropriate inferential dependencies. This is something that the Bayesian models of causal inference clearly do. It is also something that within-compound associations clearly fail to approximate.

Acknowledgments

The preparation of this article was supported by AFOSR FA 9550-08-1-0489. The authors wish to thank Betty Huang and Aaron Placensia for assistance with data collection.

References

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*.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*, 367-405.

Courville, A. C., Daw, N. D., Gordon, G. J., & Touretzky, D. S. (2003). Model uncertainty in classical conditioning. In S. Thrun, S. L, & B. Schoelkopf (Eds.), *Advances in Neural Information Processing Systems 16.* Cambridge, MA: MIT 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, *Handbook of contemporary learning theories*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

Hull, C. L. (1952). *A behavior system: An introduction to behavior theory concerning the individual organism.* New Haven, CT: Yale University Press.

Kruschke, J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning & Behavior, 36 (3)*, 210-226.

Lu, H., Rojas, R. R., Becker, T., & Yuille, A. (2008). Sequential causal learning in humans and rats. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society.*

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). San Diego, CA: Academic Press.

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

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 (3)*, 759-783.

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.