Title
Effects of Causal Strength on Learning from Biased Sequences

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 28(28)

ISSN
1069-7977

Authors
Danks, David
Schwartz, Samantha

Publication Date
2006

Peer reviewed

Effects of Causal Strength on Learning from Biased Sequences

David Danks (ddanks@cmu.edu)
Department of Philosophy, Carnegie Mellon University, 135 Baker Hall
Pittsburgh, PA 15213 USA; and
Institute for Human & Machine Cognition, 40 S. Alcaniz St.
Pensacola, FL 32502 USA

Samantha Schwartz (sschwartz@andrew.cmu.edu)
Carnegie Mellon University, 135 Baker Hall
Pittsburgh, PA 15213 USA

Abstract

Most research on step-by-step causal learning has focused on the various possible effects early correlations (in a sequence) can have on a learner's causal beliefs. Recent work has suggested that more information about an individual's learning strategy can be extracted by examining the *slope* of the learner's causal belief trajectory over time after the world changes. We examined step-by-step causal learning from biased sequences with large probabilistic dependencies, using three analyses: testing for primacy vs. recency effects; classifying learning type based on learning curve slope; and a novel analysis based on the patterns of belief change found across multiple sequences. We found few standard order effects (and all of those were primacy effects), and people seemed to be reasoning in a more "model-based" manner than had previously been demonstrated. More generally, the effects of prior observations on subsequent learning appear to be substantially subtler than previous analyses revealed.

Introduction and Related Research

Causal beliefs play a central role in many areas of cognition (Sloman, 2005), and the psychological processes governing causal learning have been the focus of substantial research. The primary psychological work on causal learning has focused on causal inference "in the long run" (Cheng, 1997; Cheng & Novick, 1992; Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks, 2004; Griffiths & Tenenbaum, 2005; Perales & Shanks, 2003; White, 2003). The resulting theories aim to explain and predict how people's causal beliefs depend on observed statistics and prior knowledge when presented with a sufficiently large number of cases.

In contrast, we focus here on the stepwise learning problem, which has received much less attention (though see, e.g., Danks, Griffiths, & Tenenbaum, 2003; Shanks, 1995; Shanks & Dickinson, 1987; and papers discussed below). The goal in this setting is to characterize the ways in which people's beliefs change upon the observation of one (or a few) cases. Thus, the resulting theories aim to predict and explain the step-by-step learning curves for sequences of cause-effect observations.

A natural experimental technique for investigating case-by-case causal belief change is the use of biased sequences: ones in which the first and second halves of the sequence exhibit significantly different correlations between the putative cause and the effect. The contrast in the statistics for the first and second halves of the sequence enable us to focus on the ways in which prior observations effect the changes in an individual's causal beliefs. To maximize the contrast between the sequence halves, we focus on conditions in which the correlation presented in the first half is exactly balanced out by the correlation of the second half. This combination results in zero correlation between the putative cause and the effect over the course of the entire sequence. Thus, any differences in final causal beliefs should be solely a result of order effects.

For sequences with this type of internal structure, there are two obvious potential order effects. Primacy effects occur when the final causal beliefs are biased towards the initial correlation (as found in Dennis & Ahn, 2001). In contrast, recency effects occur when the final causal beliefs are biased towards the second half correlation (Catena, Maldonado, & Cándido, 1998; López, Shanks, Almaraz, & Fernández, 1998; Collins & Shanks, 2002).

Two different types of theories of step-by-step causal learning have been proposed in order to account for such order effects. Associationist or error-correction models (e.g. Rescorla & Wagner, 1972; Pearce, 1994) predict that causal beliefs should change in response to the learner's prediction errors. These models thus "track" the recent correlations in the sequence, and so are invariably thought to lead to recency effects. In contrast, theories based on explicit mental models (Dennis & Ahn, 2001) hold that the learner develops an explicit model of the underlying causal relationship during the course of observations. Subsequent observations are interpreted in light of that model and, when the model is sufficiently strong, contradictory evidence is discounted (Einhorn & Hogarth, 1978; Hogarth & Einhorn, 1992). Because of this discounting, evidence in the second half of the learning sequence has less impact on the learner's causal beliefs, which is thought to result in primacy effects.

Danks & Schwartz (2005) argued on theoretical grounds that one cannot simply infer that associationist theories always predict primacy effects, while model-based theories always predict recency effects. If an error-correction model has a time-varying learning rate (which is typically necessary for convergence in the long run; see Danks, 2003), then such a model has the potential to exhibit primacy effects for certain sequences. On the other hand, theories based on explicit models can exhibit recency effects if the learner changes her mental model during the second part of the sequence (since subsequent cases would then be reinforcing the model, and so should be *over*weighted). Because there are theories of each type that predict order effects of each type, primacy/recency effects alone are insufficient to decide between associationist and model-based theories.

Instead, the shape of the causal learning curve after the midpoint—i.e., when correlations "switch"—can provide a more robust measure for classifying step-by-step learners (Danks & Schwartz, 2005). At the sequence midpoint, the learner's causal beliefs should be extremal (within the sequence). Thus, the learner should have the largest prediction errors, but also be maximally confident in her beliefs. Associationist theories, therefore, will always predict that learners will have their largest shifts in causal belief immediately after the midpoint (since errors will be largest at that point). In contrast, model-based theories predict that the largest shifts in causal belief should occur significantly *after* the switch point (namely, whenever the learner's mental model shifts in response to the evidence).

Danks & Schwartz (2005) explored order effects using a variety of sequences that varied significantly in length, though they all had the same bias within each half-sequence. They found only slight primacy effects for only a subset of sequences, which suggests a (weak) preference for model-based learning. In contrast, when they classified learners using the above criterion (time of largest belief shifts), they found primarily associationist behavior, though with a non-trivial number of (apparently) model-based learners. Moreover, for both analysis methods, there were no systematic effects of sequence length. The analysis based on learning curve shape thus revealed more about the learning process than using only final ratings.

In their experiment, they used relatively weak causal strengths compared to previous studies, such as Dennis & Ahn (2001). For example, the positive-correlation halves of the sequences had $P(E \mid C) = 0.75$ and $P(E \mid \neg C) = 0.25$, resulting in $\Delta P = 0.5$ and power PC (Cheng, 1997) causal power = 0.67. (Negative-correlation sections were exactly opposite.) Quite different order effects might occur for sequences with larger biased causal strengths. In addition, their analysis used only high-level features of the learning curve shape, and more information might be available with more sophisticated learning curve analyses. We here report the results of an experiment using sequences with a range of probabilistic dependencies, as well as a novel analysis technique for learning curves from biased sequences.

Experiment

Participants

Forty Carnegie Mellon students were compensated $10 each for participation. The experiment took approximately forty minutes to complete.

Design and Materials

The experiment was done on computers. The experiment cover story placed participants as doctors researching the causal relationships between native plants and skin diseases found on foreign islands. Over the course of the experiment participants traveled to different islands, with a new disease/plant sequence for each island.

Participants were first provided an introduction to the information they would be given, as well as to the mechanism for providing responses. Before seeing the first experimental sequence, participants were shown four cases to familiarize themselves with the experiment interface, and offered an opportunity to ask questions.

On each "island," participants interviewed forty villagers to learn about their health. For each observed case, participants were told whether or not that individual had been exposed to the local plant, and also whether that person had a specific skin rash. After each observed case, participants were asked, "How much does the plant cause the rash?" They responded using a slider that ranged from +100 (the plant "always caused" the rash) to –100 (the plant "always prevented" the rash), with 0 indicating no causal relationship. The numeric value for the slider was provided, but to avoid anchoring effects, the slider was repositioned at 0 after each response.

Each participant saw five different sequences of forty cases each. Four of the five sequences were significantly biased: the first half of the sequence had a positive (or negative) correlation between Rash and Plant, and the second half had the opposite correlation. An unbiased sequence was included as a control condition. Overall, in each individual sequence, Rash and Plant were uncorrelated, and $P(\text{Plant}) = P(\text{Rash}) = 0.5$. The precise segment statistics (and model predictions) are given in Table 1.

Table 1: Half-sequence statistics

Name of the half-sequence	*P*(Rash | Plant)	*P*(Rash | No Plant)	ΔP	Causal power
Strong –	0.1	0.9	-0.8	-0.89
Strong +	0.9	0.1	0.8	0.89
Weak –	0.3	0.7	-0.4	-0.57
Weak +	0.7	0.3	0.4	0.57
Unbiased	0.5	0.5	0.0	0.0

We will refer to the five sequences by strength, followed by '+/-' or '-/+' (when appropriate). Thus, *Strong* +/- indicates the sequence in which the participant saw the *Strong* + half-sequence, followed by the *Strong* – half-sequence.

We used four different presentation orders; in each order, the first sequence had a Strong bias, and every sequence was followed by one in which the ΔP of the first half differed by at least 0.4 (so that no sequence was followed by a "close" one). Importantly, participants were not told that any of the sequences had an internal bias, or that there might be a change at the sequence midpoint. Regardless of presentation order, every participant saw exactly the same case ordering for each sequence. Although this potentially introduces a confounding factor, it was necessary to enable any between-participant data analysis at points other than the midpoint and endpoint of a sequence.

Results and Discussion

Five individuals were removed from the data analysis due to inability to follow experimental instructions.[1] We then performed three different types of analyses.

Traditional Order Effect Analysis. Figure 1 provides the mean midpoint and final ratings for the 35 participants in each of the five sequences (error bars are 95% confidence intervals). As expected, the midpoint ratings for all of the biased sequences were highly significant (all $p < .001$; one-sample two-tailed t-test), suggesting that participants were sensitive to the probabilistic dependencies in the data. Somewhat surprisingly, the mean midpoint rating for the unbiased sequence was also significantly different from zero ($p < .05$). We currently have no explanation for this finding.

We tested for order effects in two different ways. A "normative" order effect occurs when final ratings are significantly different from zero. We found normative primacy effects in the *Strong +/-* ($p < .05$) and *Weak -/+* ($p < .01$) conditions. A "subjective" order effect occurs when an individual's ratings are significantly different for sequences with different biases. We also found subjective primacy effects for *Strong* +/- and *Weak* -/+ conditions. For each of these sequences, the mean final ratings were significantly different from the mean final ratings in the three sequences with opposite (or no) bias.[2]

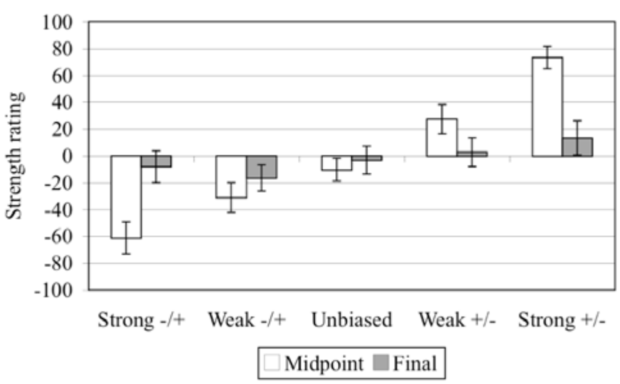

Figure 1: Midpoint and final mean ratings ($N = 35$)

These results are consistent with the findings of Danks & Schwartz (2005) for weak causal strengths: order effects were found in only some conditions, and all order effects were primacy effects. We also did a between-group comparison of the mean midpoint and final ratings of (i) the Weak conditions in the present experiment, and (ii) the 32- and 48-case sequences of Danks & Schwartz (2005). The only significant difference was between the midpoint ratings of the *Weak* +/- and 48-case sequences ($p < .05$). Although such comparisons are notoriously problematic, this analysis suggests that the present experiment (partially) replicates the results of Danks & Schwartz (2005).

Learning Curve Classification. The classification of individuals into distinct "learning types" based on the high-level shape of the learning curve reveals a similar, but not identical, picture to that found in Danks & Schwartz (2005). The basic classification method for this analysis compares (for each individual) the changes in rating between (a) the midpoint and ¾-point; and (b) the ¾-point and final rating. If change (a) is larger than change (b), then the individual is learning as if "Associationist"; if change (b) is larger, then the individual's learning is as if "Model"-based.

To avoid inferences based on insignificant differences, we classified any individual whose changes (a) and (b) were within two points of one another as "Indeterminate." Similar results (though with a corresponding increase in the number of "Indeterminate" individuals) were obtained for larger thresholds. In addition, qualitatively similar results were obtained when the learning curves were "smoothed" in various ways (e.g., using the mean ratings in some window around the point, rather than the point itself). Table 2 provides the classification of all 35 individuals in each of the five conditions.

Table 2: High-level classification of individuals

	Associationist	Model	Indeterminate
Strong -/+	23	6	6
Weak -/+	22	2	11
Unbiased	18	7	10
Weak +/-	11	13	11
Strong +/-	19	10	6

As one would expect, there are many more "Indeterminate" individuals in the Weak and Unbiased conditions than in the Strong conditions. Those sequences induced much weaker causal beliefs in the participants at the midpoints, and the subsequent changes were correspondingly smaller as well. Thus, it is more likely that the relevant shifts in any particular individual's learning curves in those conditions will be (approximately) equal.

Overall, substantially more individuals are classified as Associationist learners in the -/+ sequences than in the -/+ sequences. In contrast, Danks & Schwartz (2005) did not find a substantial difference between the classifications for +/- and -/+ sequences of similar length. However, since this classification criterion is relatively coarse (and the 2005 experiment had relatively small sample sizes), we do not place substantial weight on this difference.

[1] Specifically, four individuals responded with the impact of each particular case (rather than integrating over the cases that they had observed). One individual gave only zeros for ratings.

[2] *Weak* -/+ vs. *Strong* +/-: $p < .01$; *Weak* -/+ vs. *Weak* +/-: $p < .02$; all other relevant pairs: $p < .05$ (all two-sample paired t-tests)

As a check on this analysis method, we performed the same analysis on a uniform population of associationist learners. Specifically, we simulated 35 individuals, each of whom learned using the augmented Rescorla-Wagner model (Van Hamme & Wasserman, 1994) with individual-specific parameter values.[3] Since we used fixed sequences in the experiment, we were able to calculate a precise causal belief learning curve for each individual. As one would expect, the simulated learning curves were qualitatively similar, though with minor differences due to variations in parameter values. These simulated individuals have no noise in their responses; their ratings exactly correspond to their current beliefs. To better approximate realistic behavior, we did 1000 runs in which we applied Gaussian noise (mean=0, sd=5) to the ratings for each individual.

We classified these 35,000 (noisy) learning trajectories using the above method. The classification profile (in thousands) is given in Table 3. The simulation classification finds more Associationist behavior in the -/+ sequences than in the +/- sequences, just as in the empirical classification. Thus, it seems reasonable to conclude that there are a non-trivial number of associationist experimental participants. At the same time, there are notable differences between the two classifications, particularly many fewer Indeterminate individuals in the simulation classification (especially for *Weak* -/+). Thus, we suspect that we have a mixed population of causal learners (see also Lober & Shanks, 2000). Unfortunately, we do not know of any proposed computational theories of step-by-step model-based learning, and so we cannot produce a complementary set of simulated model-based learners.

Table 3: Classification of noisy simulated individuals

	Associationist	Model	Indeterminate
Strong -/+	34.2	0.4	0.4
Weak -/+	31.9	1.7	1.4
Unbiased	19.9	7.0	8.1
Weak +/-	16.4	14.1	4.5
Strong +/-	23.3	8.0	3.7

Comparisons of Belief Change Patterns. Our final set of analyses focused on a previously unexplored feature of this type of data. By design, for each bias, the first half of the +/- sequence is the same as the second half of the corresponding -/+ sequence (and *vice versa*). Thus, participants see the exact same sequence of 20 cases twice: once as the first half of a sequence, and once as the second half. We can therefore analyze order effects on learning by comparing *directly* the changes in their ratings when presented with the same sequence of cases in different settings (either the start or the midpoint of a sequence). Any significant differences between the changes that participants made on some particular case are presumably due to order effects.

[3] λ = 1.0; all other parameters drawn uniformly from: $\alpha_B \in [0.6, 0.8]$; $\alpha_C \in [0.7, 0.9]$; $\alpha_{\neg C} \in [-0.3, -0.4]$; $\beta_1, \beta_2 \in [0.1, 0.2]$.

We have four different biased half-sequences: *Strong* +, *Strong* –, *Weak* +, and *Weak* –. Figures 2-5 show the mean case-by-case *changes* for each of the half-sequences in each of the relevant full sequences. (Note that the y-axis scale is not the same in the four figures.) For example, the *Strong* – graph (Figure 2) compares the first half of the *Strong* -/+ sequence with the second half of the *Strong* +/- sequence.

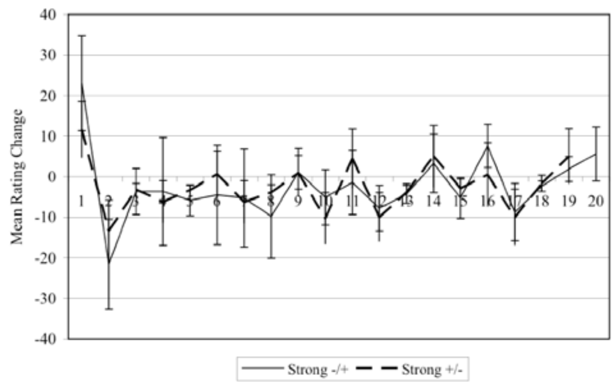

Figure 2: Mean changes in *Strong* – half-sequence

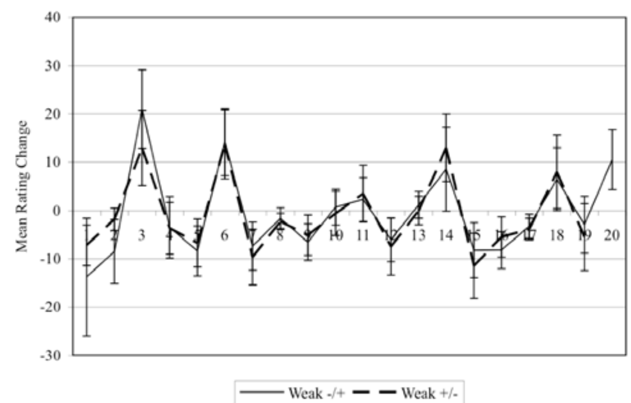

Figure 3: Mean changes in *Weak* – half-sequence

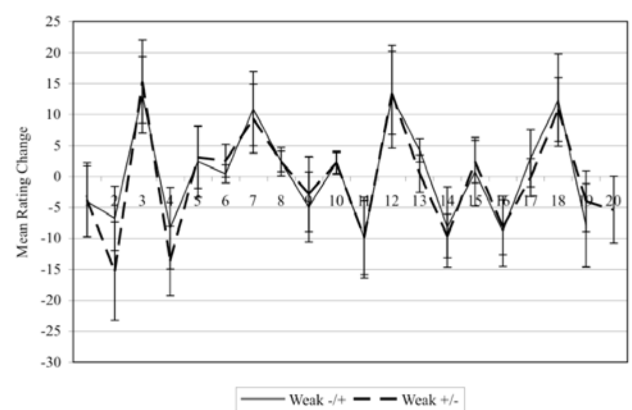

Figure 4: Mean changes in *Weak* + half-sequence

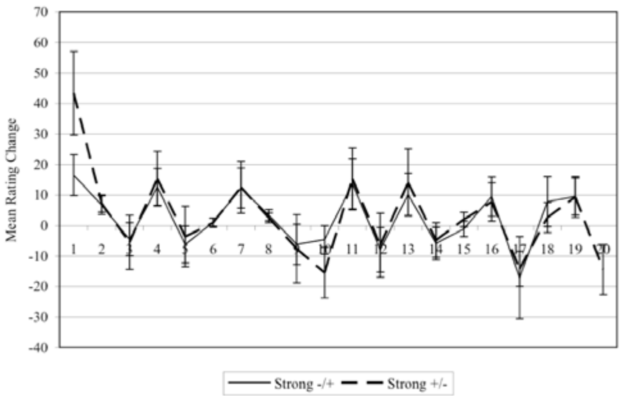

Figure 5: Mean changes in *Strong* + half-sequence

We then tested for significant differences in each pair of 19 case-to-case changes in each of the four half-sequences. For example, one test is for differences between (i) the change from the first to second case in the *Strong* -/+ sequence; and (ii) the change from the 21st to 22nd case in the *Strong* +/- sequence (point 1 in Figure 2). Out of the 76 distinct tests (all two-sample paired t-tests), there were only seven significant differences (shown in Table 4).

Table 4: Significant differences between participant changes

Half-sequence	**Location of Significant Difference (1-20)**	***p*-value**
Strong –	NONE	
Weak –	3	< 0.05
Weak +	2	< 0.05
	4	< 0.05
	13	< 0.05
	19	< 0.05
Strong +	1	< 0.01
	10	< 0.01

Note that there were no significant differences between the changes in the two *Strong* – half-sequences. In other words, over the whole population, participants' responses for particular cases in the strongly negatively biased half-sequence were *not* sensitive to the "context": whether a case occurred in the first or second half was not (in general) a significant factor in predicting participants' responses to it.

As a comparison, we returned to the 35 simulated associationist learners (× 1000 noise applications). The noise applications now enable us to test the stability of significant differences between pairs of changes. Changes that are robustly significantly different should be different for most of the noise applications. We thus tested all pairs of changes for significant difference (defined to be $p < .05$ on the two-sample paired t-test). The number of significant differences for each pair of changes in those 1000 runs provides an estimate of the "robustness" of the difference.

Every pair of changes was significantly different at least 40 times, but only one pair was significantly different in all 1000 runs. In fact, only ten pairs of changes were significantly different even 30% of the time. Those pairs of changes are listed in Table 5 (with percent of runs).

Table 5: Significant differences between simulated changes

Half-sequence	**Location of Significant Difference (1-20)**	**Percent of Runs**
Strong –	1	100.0
	4	49.2
	5	42.2
	7	34.2
Weak –	1	71.4
	2	62.6
	3	98.5
	6	78.8
Weak +	3	59.4
Strong +	1	33.4

There are two salient differences between the simulation analysis and the empirical data, and in both cases, the performance of the simulated learners is readily explained. First, all robustly significantly different pairs in the simulated data occurred in the first third of the half-sequence, while the significant differences in the empirical data occurred throughout the half-sequences. This biased distribution of significant differences is entirely to be expected for the simulated data. Associationist learners should have large shifts immediately after the midpoint in a sequence compared to the shifts at the start of a sequence. Thus, we should find significant differences between the early changes depending on whether they came at the start or just after the midpoint. Moreover, since the associationist learner rapidly converges within each half-sequence, there should be very few context effects for the later portions of the half-sequence. This explanation is clearly closely related to the features of associationist learning that justified our second analysis (using change sizes after the midpoint).

Second, the robustly significantly different pairs in the simulation clustered in the negative half-sequences, whereas the significantly different pairs of changes in the empirical data tended to occur in the positive half-sequences. But the previous asymmetry (between changes at the start and at the midpoint) should be heightened for negative half-sequences, since associationist learners have almost no changes in belief *for the target cause* when presented with a negative correlation.

Overall, this analysis thus supports our prior (tentative) conclusion that our empirical population is almost certainly not composed entirely of associationist learners.

Conclusion

A powerful experimental tool for discerning the case-by-case manner in which learners change their causal beliefs is the use of biased sequences: those in which the correlation between the putative cause and the effect shifts over the course of learning. We used multiple sequences of this type to help determine whether causal learning is principally based on associationist/error-correction methods, or on explicit models of the underlying causal structure.

We found that the standard experimental focus on only primacy vs. recency effects fails to capture the subtlety of order effects on causal learning. In some conditions, the internal bias of a sequence does seem to produce primacy effects in an individual's final causal beliefs. However, the story is more complicated than just "first half correlations matter more than second half correlations" (or *vice versa*).

By looking at both the overall shape of the causal learning curve and the precise pattern of changes (in response to identical data), we found that previously observed cases do not seem to exert a uniform influence on all subsequent cases. This focus on the step-by-step changes enables us to analyze the participant data in substantially more detail than simply looking at mean midpoint and final ratings. In particular, any significant differences in changes for the same cases in the same half-sequence (but in two different full sequences) are almost certainly due to order effects, and so we can make more direct inferences than were possible using only mean final ratings.

In addition to pointing towards the subtle nature of order effects, these analyses also strongly suggest that our participant population was not uniform with regards to causal learning strategy (see also Lober & Shanks, 2000). We aim in future experiments to examine more carefully the individual differences in learning strategy.

Acknowledgments

D. Danks was partially supported by grants from the Office of Naval Research, and the National Aeronautics and Space Administration.

References

Catena, A., Maldonado, A., & Cándido, A. (1998). The effect of the frequency of judgment and the type of trials on covariation learning. *Journal of Experimental Psychology: Human Perception and Performance, 24*, 481-495.

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

Cheng, P.W., & Novick, L.R. (1992). Covariation in natural causal induction. *Psychological Review, 99*, 365-382.

Collins, D. J., & Shanks, D. R. (2002). Momentary and integrative response strategies in causal judgment. *Memory and Cognition, 30*, 1138-1147.

Danks, D. (2003). Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology, 47*, 109-121.

Danks, D., Griffiths, T. L., & Tenenbaum, J. B. (2003). Dynamical causal learning. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15*. Cambridge, Mass.: MIT Press.

Danks, D., & Schwartz, S. (2005). Causal learning from biased sequences. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Dennis, M. J., & Ahn, W. (2001). Primacy in causal strength judgments: The effect of initial evidence for generative versus inhibitory relationships. *Memory and Cognition, 29*, 152-164.

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review, 85*, 396-416.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111*, 3-32.

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

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24*, 1-55.

Lober, K., & Shanks, D. R. (2000). Is causal induction based on causal power? Critique of Cheng (1997). *Psychological Review, 107*, 195-212.

López, F. J., Shanks, D. R., Almaraz, J., & Fernandez, P. (1998). Effects of trial order on contingency judgments: A comparison of associative and probabilistic contrast accounts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 672-694.

Pearce, J. M. (1994). Similarity and discrimination: A selective review and a connectionist model. *Psychological Review, 101*, 587-607.

Perales, J. C., & Shanks, D. R. (2003). Normative and descriptive accounts of the influence of power and contingency on causal judgement. *The Quarterly Journal of Experimental Psychology, 56A*, 977-1007.

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*. New York: Appleton-Century-Crofts.

Shanks, D. R. (1995). Is human learning rational? *The Quarterly Journal of Experimental Psychology, 48A*, 257-279.

Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgment. In G. Bower (Ed.), *The Psychology of Learning and Motivation, vol. 21*. San Diego: Academic Press.

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives*. Oxford: Oxford University Press.

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.

White, P. A. (2003). Making causal judgments from the proportion of confirming instances: The pCI rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 710-727.