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* (pp. 542-547). Mahwah, N.J.: Lawrence Erlbaum Associates.

# Causal 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**

Multiple psychological theories of causal learning provide case-by-case updating rules: given my current causal beliefs about the world and a novel case, how should I change those beliefs? Most of these theories predict some type of order effect: biased and unbiased sequences of cases will lead to different final causal beliefs, even if the overall statistics are identical. This paper describes an experiment that (i) finds only small order effects that (ii) are not dependent on the number of observed cases, and in which (iii) observed patterns of belief *change* during the sequences are not explained by various proposed algorithmic theories.

**Keywords:** Causal models; biased observations; learning.

## Introduction

Causal knowledge and beliefs play a significant role in much of our everyday cognition. A range of theories have been proposed over the past fifteen years to explain human causal learning, and in particular, learning from sequences of observations or manipulations of the world. That is, they predict the inference of causal relationships from a sequence of individual cases, each of which is either observed or produced by the learner.

Algorithmic theories offer explicit, case-by-case updating rules. Most notably, this group of theories includes standard associationist models (e.g., Pearce, 1994; Rescorla & Wagner, 1972). In contrast, computational theories aim to predict our stable, long-run causal beliefs. These theories range from the purely probabilistic (e.g., conditional $\Delta P$ model of Cheng & Novick, 1992; Spellman, 1996), to theories with more robust metaphysics based on observed probabilities. This last group includes both Cheng's (1997) power PC theory, and various theories based on causal Bayesian network structure inference (a partial list includes Danks, Griffiths, & Tenenbaum, 2003; Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks, 2004; Griffiths & Tenenbaum, in press; Lagnado & Sloman, 2004; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003; Tenenbaum & Griffiths, 2001; Waldmann & Martignon, 1998).

These types of theories are connected in at least two ways. First, the long-run behaviors of many algorithmic theories are characterized by (independently proposed) computational theories. Second, many of these theories—both algorithmic and computational—correspond to maximum likelihood estimates of parameters in specific Bayesian network structures. An overview of these connections can be found in Danks (in press).

In this paper, we focus on a salient feature of essentially all of the algorithmic theories (that are currently considered viable). They all predict some sensitivity to the order of case presentation: different orderings of the same set of cases will (sometimes) lead to different responses at the end of the sequence. Various theories predict different order effects, and so biased sequences have been used to test the various algorithmic-level theories.

In contrast, essentially all of the standard computational-level theories, including power PC, conditional $\Delta P$, and standard Bayesian network learning (whether Bayesian updating or constraint-based) assume that the observed cases are independently distributed. That is, they assume that the probability of observing a case does not depend on the previous trial(s). As a result, they make no prediction about order effects. This does *not* mean that they explicitly predict the *absence* of order effects. Rather, since a basic assumption is violated (e.g., if a sequence is biased in certain ways), these theories do not make any clear, determinate predictions.

## Primacy vs. Recency

Consider the following three distinct types of correlations within some sequence of observed cases:

- Pos/Neg: The first half of the sequence has a positive correlation between variables *C* and *E*, and the second half has a negative correlation;
- Neg/Pos: The first half has a negative correlation, and the second half has a positive correlation; and
- Even: *C* and *E* are uncorrelated during the sequence.

There are two natural types of order effects. A *primacy effect* occurs if the initial cases have a greater weight in the inference process than later ones. In that case, the first sequence should result in a positive perceived causal strength, the second in a negative causal strength, and the third with a zero causal strength.

In contrast, a *recency effect* occurs when the later cases have a greater impact on perceived causal strength than earlier ones. The response profile of a recency effect in

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* (pp. 542-547). Mahwah, N.J.: Lawrence Erlbaum Associates.

these situations is: negative perceived causal strength in the first sequence, positive strength in the second sequence, and zero causal strength in the third.

Four recent studies have tried to determine whether people's causal learning is subject to primacy or recency effects. López, Shanks, Almaraz, & Fernández (1998) and Collins & Shanks (2002) both found that people exhibit recency effects. In contrast, Dennis & Ahn (2001) and Marsh & Ahn (under review) have found evidence of primacy effects.

López, *et al*. (1998) placed significant memory demands on their experimental participants, and experiments in Marsh & Ahn (under review) strongly suggest that those memory demands are, at least in part, responsible for López, *et al*. finding recency effects. More significantly, Collins & Shanks (2002) found that increasing the frequency of judgments from only at the end-of-sequence to every ten trials increases the likelihood of finding a recency effect (see also Catena, Maldonado, & Cándido, 1998). Beyond this effect, the interaction between judgment frequency and size of primacy/recency effects is poorly understood.

In addition, a plausible factor in the occurrence of a primacy vs. recency effect is the number of trials seen. Despite this, only three different numbers of cases have been used, and all three have been quite long (40, 80, and 160 cases). Dennis & Ahn (2001) found comparable primacy effects for 40 and 80 case sequences. The dependence of primacy/recency effects (and sizes) on the number of cases seen has not otherwise been studied.

## Two Theoretical Explanations

Essentially two types of theoretical explanations have been offered to explain the occurrence of primacy and recency effects: associative theories, and explicit model-based theories. In general, the former have been offered to explain recency effects, and the latter to explain primacy effects. However, each type of theory can actually explain both effects, depending on particular (untested) assumptions.

López, *et al*. (1998) and Collins & Shanks (2002) both advocate associative learning theories, though of different types: Pearce's (1994) configural cue associationism, and a mix of associationism and between-judgment adjustments, respectively. At a high level, associationist theories have a set of associative (causal) strengths that are adjusted by error-correction. That is, if $V_j$ is the associative strength of $j$ (possibly a configural cue), then after seeing a new case, we change $V_j$ by:

$$\Delta V_j = \mathit{Rate} \times (\mathit{Actual} - \mathit{Prediction}).$$

Various associative theories are distinguished by the prediction function, the encoding of the actual event, and the rate parameter. For example, the Rescorla-Wagner (1972) model has a constant rate parameter, represents the actual event by a binary variable, and generates predictions by summing the current associative strengths of the cues that occur in a particular case.

Associative theories are typically thought to produce recency effects. Because the models are error-driven, they essentially try to track the "current" state of the world. As a result, an associative model presented with the Pos/Neg sequence should (if the sequence is long enough) "learn" the Neg distribution.

Other types of associative models can also produce a primacy effect, though they have not been explicitly advocated in the causal learning literature. In most standard associative models (e.g., Rescorla-Wagner and variants, Pearce), the rate parameter is constant. As a result, these models do not converge for many situations to any asymptote, but only to a distribution of values (Danks, 2003; Yuille, 2005). A natural adjustment to an associative model is to allow for time-varying rate parameters, and in particular, rate parameters that grow smaller with time. Most such models will have well-defined asymptotes, rather than equilibrium distributions. More importantly, many such models will exhibit primacy effects (depending on the time variation and sequence length).[1]

The second type of theoretical explanation offered for primacy and recency effects is explicitly model-based (Dennis & Ahn, 2001; Marsh & Ahn, under review). The central intuition behind these theories is that learners develop an explicit model of the causal structure of the situation based on initial evidence, and then interpret subsequent observations in light of that model. In contrast with associative theories, explicit model-based models have rarely been computationally fully-specified.

Explicit model-based theories were introduced by Ahn and her colleagues to explain primacy effects. Based on the first few observed cases in a sequence, the learner converges on an hypothesis about the causal structure underlying the cases. Subsequent observations are then interpreted in light of that model. In particular, evidence supporting the hypothesis is weighted more heavily than evidence contradicting the hypothesis (see also Einhorn & Hogarth, 1978; Hogarth & Einhorn, 1992; Klayman & Ha, 1987). So in structured sequences, the initial observations dominate the later ones, resulting in a primacy effect.

Alternately, explicit model-based theories can also predict a recency effect, depending on the rate at which a model is learned or changed. Any such theory must allow for the possibility that the learner changes her mind after sufficient counter-evidence. And if such a change occurs before the end of the sequence, then the remaining cases should receive substantially more weight than the initial cases (since they now support the learner's explicit model). Thus, if the evidence in the second half of the sequence prompts the learner to adjust her explicit model early, then we would expect to see (at least some) evidence for a recency effect.

Since both theory-types can sometimes predict a primacy effect and sometimes a recency effect, we should find a different behavioral measure to separate them. One natural

[1] Note that this adjustment does not model learner fatigue, but rather simple discounting of evidence based on the number of previously observed cases. Thus, Dennis & Ahn's (2001) finding of primacy effects even after an explicit attempt to control participant fatigue does not falsify these models.

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* (pp. 542-547). Mahwah, N.J.: Lawrence Erlbaum Associates.

candidate measure is the shape of the learning curve at the midpoint of learning. If the sequence is biased so that the first half shows one correlation and the second half shows the opposite, then the cases immediately after the midpoint would involve the largest prediction errors, but also presumably the greatest confidence in an explicit model.

Therefore, essentially all associative theories, even those with (intuitively natural) time-varying rate parameters, predict that the change in associative strength during the cases immediately after the midpoint should be larger than at any other point in the second half of the sequence. The learner is making more errors at this point than at any other time in the second half, and so should be changing her opinion more than at any other time.

In contrast, essentially all explicit model-based theories should predict that the change in perceived causal strength during these cases should be smaller than at any other time in the second half (or at least, there should be another period in which the change is much larger than immediately after the midpoint). Since the learner's confidence in her model is greatest, she should most discount conflicting observations. Substantial change in perceived causal strength should occur only after she begins to doubt her explicit model.

## Experiment

The current state of experimental results and theoretical explanations points to two natural questions:

1. How does the size of the order effects (if any) depend on the number of cases observed? (Note that a tentative answer to this question also speaks to the overall debate about the prevalence of primacy and recency effects.)
2. Based on ratings immediately following the midpoint, does causal learning appear to be associationist, explicitly model-based, or something else?

To find (partial) answers to these questions, we presented experimental participants with Pos/Neg, Neg/Pos, and Even sequences of cases.

### Participants

51 Carnegie Mellon University students volunteered to participate and were compensated $10. The experiment took approximately 40 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 given an introduction explaining what information would be given, as well as how they were to provide their responses. Before seeing any actual cases, participants were shown a brief sequence to familiarize themselves with the experiment interface, and offered an opportunity to ask questions.

On each island, participants interviewed varying numbers of individual villagers to learn about their health. For each observed case, participants were told whether or not that individual had been exposed to the native plant, and if 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 prevented" the rash) to +100 (the plant "always caused" the rash), with 0 indicating no causal relationship. The numeric value for the slider position was also provided. To avoid anchoring effects, the slider was repositioned at 0 after each rating.

There are four combinations of plant/rash values. In every sequence, there were an equal number of cases of all four types, resulting in zero correlation (and $P$(Plant) = $P$(Rash) = 0.5). In the biased sections of Pos/Neg and Neg/Pos sequences, the conditional probabilities were:

*Pos*: $P$(Rash | Plant) = 0.75 and $P$(Rash | No plant) = 0.25

*Neg*: $P$(Rash | Plant) = 0.25 and $P$(Rash | No plant) = 0.75

(And so we have $\Delta P$ = .5 and -.5, causal power = ⅔ and -⅔.)

Participants saw six sequences of cases in total. After the first three sequences, they solved several distractor math problems. The order of sequence lengths were fixed for all participants: 8, 80, 8, 32, 16, and 48 cases. This ordering aimed to minimize fatigue effects by balancing the number of cases before and after the distractor task. In each group of three trials, participants saw one Pos/Neg sequence, one Neg/Pos sequence, and one Even sequence. The pairing of sequence type and sequence length was randomized across participants. For every type-length pair, a fixed sequence of cases was used across participants.

### Results and Discussion

Figure 1 presents the mean causal ratings at the sequence midpoints (error bars indicate standard error). All pairwise (two-tailed $t$-test) comparisons within each sequence length were significantly different ($p < .02$), and 60% were highly significant ($p < .001$).[2] Participants were responsive to the cause-effect correlations, and not just responding randomly.

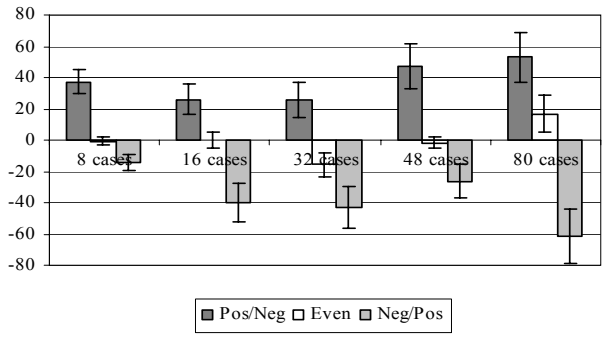


Figure 1: Mean midpoint ratings

[2] The significance levels were: $p < .001$: P/N vs. N/P (8, 16, 32, 48, 80), P/N vs. E (8, 48), N/P vs. E (16, 80); $p < .01$: P/N vs. E (16, 32), N/P vs. E (8); $p < .02$: P/N vs. E (80), N/P vs. E (32, 48)

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* (pp. 542-547). Mahwah, N.J.: Lawrence Erlbaum Associates.

The first question motivating this experiment focused on whether any order effects occur, and if so, whether they are dependent on sequence length. Figure 2 provides the mean final ratings. There were only slight primacy effects. Only five of the fifteen mean ratings were significantly different from zero, and none of them were highly significantly different.[3] Moreover, in the three unbalanced conditions with mean final ratings significantly different from zero, the order effect was always primacy. There was no pattern to the conditions in which order effects occurred, suggesting that sequence length is not an important factor.

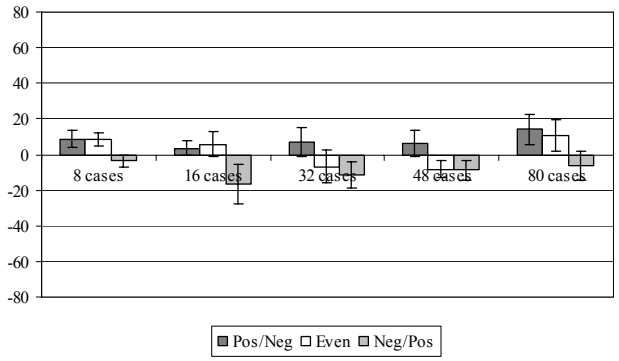


Figure 2: Mean final ratings

The second question focuses on within-subject patterns of belief change after the midpoint. There are multiple plausible associationist and explicit-model learning algorithms, and participants saw different sequence lengths. Thus, we aimed for a classification scheme that did not depend on those details. In particular, for some sequence of ratings, we operationalized the participant-types as:

*Associationist*: The magnitude of change between the midpoint and ¾-point is greater than the magnitude of change between the ¾-point and the final rating.

*Explicit-Model*: The ordering of change magnitudes is reversed from the associationist.

Close examination of the data also revealed that a subset of ratings revealed very little change throughout the course of an entire sequence. We thus defined the additional type:

*Static Belief*: The absolute value of the ratings is always strictly less than some fixed threshold. (Classification into this type supersedes the previous two types.)

Table 1 gives the results of classifying every sequence ($N$ = 34 for 8-case; $N$ = 17 for others) of participant ratings, using a threshold of 10.0 for *Static* and indicating (non-*Static*) participants with equal changes by 'None'.[4] Some care must be exercised in considering the results for the 8-case sequences, since there is only one rating between the midpoint and ¾-point, and the ¾-point and final rating.

Table 1: Participant classifications

| Sequence | Type | 8 | 16 | 32 | 48 | 80 |
|---|---|---|---|---|---|---|
| Pos/Neg | Assoc. | 27 | 9 | 8 | 9 | 11 |
| | Model | 1 | 2 | 5 | 7 | 4 |
| | Static | 4 | 5 | 3 | 1 | 2 |
| | None | 2 | 1 | 1 | 0 | 0 |
| Neg/Pos | Assoc. | 14 | 12 | 11 | 8 | 7 |
| | Model | 9 | 2 | 5 | 6 | 9 |
| | Static | 10 | 2 | 1 | 2 | 0 |
| | None | 1 | 1 | 0 | 1 | 1 |
| Even | Assoc. | 9 | 7 | 9 | 8 | 5 |
| | Model | 13 | 1 | 5 | 3 | 8 |
| | Static | 8 | 8 | 1 | 3 | 2 |
| | None | 4 | 1 | 2 | 3 | 2 |

Even at this qualitative level of analysis, there consistently seems to be a distribution over the strategy-types.

In addition, participants seem to respond differently to the Pos/Neg and Neg/Pos sequences. There is a consistent bias in favor of associationist learning for Pos/Neg sequences; in the Neg/Pos sequences, there seems to be a shift towards model-based learning as the sequence length increases. At least intuitively, these two types of sequences are different. Early negative evidence is ambiguous between "no effect" and "preventive effect"; in contrast, positive evidence is almost always interpreted as confirming a generative effect. This difference is reflected in, for example, the initial bump above zero that associationist models exhibit when presented with sequences in which there is zero correlation. Thus, we conjecture that the differential behavior is a product of differences in the sequences, and not an artifact of our experimental method.

We wanted to confirm that this classification method was not artificially creating a distribution where one did not exist, and that the difference between sequence types was a meaningful one. We thus simulated 1000 individuals with augmented Rescorla-Wagner models (Van Hamme & Wasserman, 1994) with random parameter values,[5] and presented them with the fifteen possible sequences.

The classification of their rating sequences is given in Table 2 (excluding rows when no individuals were classified as using that strategy). Note that the model behaves differently on the Pos/Neg and Neg/Pos sequences, suggesting that the different behaviors in the two conditions are due (at least in part) to differences in the sequences themselves. Also, the apparent strategy distribution in the Even sequences is an artifact of the model's relatively stable behavior on long, unbiased sequences; the ratings simply

[3] The five sequences and significance levels were: $p < .02$: E (8); $p < .05$: P/N (8); $p < .10$: P/N (80); E (48); N/P (32).

[4] The results of the analysis were qualitatively similar for several different thresholds, and if we smoothed the rating sequences by defining a point's "value" by either (a) the mean, or (b) largest magnitude value for that point and the preceding three points.

[5] $\lambda = 1.0$; all other parameters drawn uniformly from: background salience $\in [0.6, 0.8]$; present cue salience $\in [0.7, 0.9]$; absent cue salience $\in [-0.3, -0.4]$; effect present/absent rate $\in [0.1, 0.2]$

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* (pp. 542-547). Mahwah, N.J.: Lawrence Erlbaum Associates.

have almost no change after the midpoint, and so (depending on exact parameter values) the changes might be slightly more or slightly less in the relevant quarters of the rating sequence.

Table 2: Simulation classifications

| Sequence | Type | 8 | 16 | 32 | 48 | 80 |
|---|---|---|---|---|---|---|
| Pos/Neg | Assoc. | 1000 | 438 | 1000 | 990 | 1000 |
| | Model | 0 | 562 | 0 | 10 | 0 |
| Neg/Pos | Assoc. | 988 | 1000 | 905 | 1000 | 1000 |
| | Model | 0 | 0 | 95 | 0 | 0 |
| | Stable | 12 | 0 | 0 | 0 | 0 |
| Even | Assoc. | 7 | 338 | 557 | 249 | 761 |
| | Model | 733 | 516 | 443 | 751 | 239 |
| | Static | 260 | 146 | 0 | 0 | 0 |

## A Different Possibility

In this section, we briefly outline a model that incorporates elements of both associationist and explicit model-based reasoning. As a result, the model straightforwardly predicts the appearance of a strategy distribution, even if all individuals are using the same model (but with different parameter values). Although the experimental data provided in this paper do not provide clear evidence in favor of this model, it is valuable to see a model that can produce a wide range of learning trajectories.

Recall that the general form of associationist models is:

$$\Delta V_j = Rate \times (Actual - Prediction).$$

The rate parameter is almost always assumed to be fixed, or in rare occasions, a monotonic function of sample size. Instead, suppose the rate parameter is a function of the learner's confidence in the current estimates (where "confidence" must be spelled out in significantly more detail, but need not be a monotonic function of sample size). We call such a model a 'confidence-based error-correction model.' The most natural function is a soft threshold: the learning rate (i.e., the ability to change one's mind) is high until the learner's confidence crosses a (soft) threshold, at which point the learning rate drops significantly.

The behavior of such models can be highly dependent on the sequence length. For relatively short sequences, this model would appear to be associationist. The high learning rate would lead the learner to be quite sensitive to changes in the system throughout the sequence. In contrast, for longer sequences (i.e., when the confidence crosses the soft threshold), the model would behave as an explicit-model theory. Because the learning rate is much lower, the learner will not significantly change her beliefs until she has observed enough cases to push her confidence back down below the soft threshold.

These models have many of the virtues of associationist models, such as relatively low memory and computational burdens. At the same time, they provide some of the benefits of explicit-model theories, such as relative stability and conscious access/control when beliefs stabilize.

To our knowledge, computational models of this type have not previously been proposed for causal learning. In order to give more substance to this high-level, rather vague description, we offer some (tentative) details about one implementation of this model-type. Denote the current causal strength estimates of the potential cause $C$ and the always-present background $B$ by $V_C$ and $V_B$, respectively, and the current confidence in those estimates by *Con*. Given a new observation, update the strength estimates by:

$$\Delta V_i = f(Con) \times \left[\delta(E)\lambda - (V_B + \delta(C)V_C(1 - V_B))\right],$$

where $\delta(X)$ is the Kronecker delta function (1 if $X$ is present, 0 if $X$ is absent), and $f$ is some function of the learner's confidence. This model uses the noisy-OR prediction function, whose stable equilibrium points are the power PC causal power predictions (Danks, *et al.*, 2003).

Obviously, the keys to this model are computation of *Con* and the function transforming that to a learning rate. One natural possibility is to let *Con* equal one minus the average "perceived" prediction error (i.e., the part in the square brackets above) over the previous $k$ cases, where the perceived error is equal to the actual error if the learner is not confident, and some fraction $\rho$ of the actual error if the learner is confident. As a measure of the current confidence, we simply use a hard threshold $\tau$ on the current average perceived prediction error. If my perceived prediction error over the previous $k$ cases is less than $\tau$, then I am confident that I am right (and so discount the current prediction error by $\rho$ in subsequent computations of averages). We then use $f(Con) = (1 - Con)$.

A range of informal simulations on biased sequences of different lengths confirms that a confidence-based error-correction model can exhibit both associationist and explicit model-based behavior. We do not intend to suggest that these simulations are definitive in any way. Rather, they are intended as proofs-of-concept that confidence-based error-correction models can generate the types of strategy shifts observed in this experiment.

## Conclusion

In summary, we found only small primacy effects in this experiment. Moreover, the slight effects did not exhibit any systematic dependence on sequence length. The relative lack of order effects, regardless of sequence length, is perhaps due to the relatively weaker causal relationships used in this experiment. Both Dennis & Ahn (2001) and Collins & Shanks (2002) used much stronger causal relationships in the biased sequences ($\Delta P$ = .8/-.8; causal power = .89/-.89), and order effects might occur only with strong causal relationships. For example, in an explicit model-based theory, people might only represent a causal relationship with a conscious model when it is particularly strong. We are currently conducting an experiment that systematically varies the strength of causal relationships within biased sequences to test this hypothesis.

The relative lack of substantial order effects is somewhat surprising in light of Collins & Shanks's (2002) finding that

increasing judgment frequency leads to greater recency effects. We obtained judgments after every trial, and so one might have expected us to find substantial recency effects, which did not occur in our data. Of course, this experiment does not constitute counter-evidence to Collins & Shanks's hypothesis, as we did not systematically manipulate judgment frequency. It does, however, suggest that other factors might explain their findings of recency effects (see also Marsh & Ahn, under review).

With regards to learning strategy, the analysis is much more challenging. Because each participant brings his or her own biases or parameters to the experiment, we should expect a greater diversity of learning curves than final ratings. However, we feel that this modeling of individual learning curves will prove central to understanding causal learning. The relatively minimal analysis provided here shows evidence for forms of both associationist and explicit model based learning. The confidence-based error-correction models described here offer one explanation for this apparent distribution of learning strategies.

## Acknowledgments

Thanks to three anonymous reviewers from the Cognitive Science conference. D. Danks was partially supported by supported by grants from the National Aeronautics and Space Administration, and the Office of Naval Research.

## 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. (in press). Causal learning from observations and manipulations. In M. Lovett & P. Shah (Eds.), *Thinking with Data*. Mahwah, N.J.: Lawrence Erlbaum Associates.

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.

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. (in press). Elemental causal induction. *Cognitive Psychology*.

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

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94*, 221-228.

Lagnado, D. A., & Sloman, S. A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 30*, 856-876.

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.

Marsh, J. K., & Ahn, W. (under review). Order effects in contingency learning: The role of task complexity. Manuscript currently under review.

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

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.

Spellman, B. A. (1996). Conditionalizing causality. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *Causal learning: The psychology of learning and motivation, vol. 34*. San Diego, Calif.: Academic Press.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science, 27*, 453-489.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Structure learning in human causal induction. In T. Leen, T. Deitterich, & V. Tresp (Eds.), *Advances in Neural Information Processing Systems 13*. Cambridge, Mass.: The MIT 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.

Waldmann, M. R., & Martignon, L. (1998). A Bayesian network model of causal learning. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Yuille, A. (2005). The Rescorla-Wagner algorithm and maximum likelihood estimation of causal parameters. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 17*. Cambridge, MA: The MIT Press.