Causal Learning With Local Computations

Philip M. Fernbach and Steven A. Sloman
Brown University

The authors proposed and tested a psychological theory of causal structure learning based on *local computations*. Local computations simplify complex learning problems via cues available on individual trials to update a single causal structure hypothesis. Structural inferences from local computations make minimal demands on memory, require relatively small amounts of data, and need not respect normative prescriptions as inferences that are principled locally may violate those principles when combined. Over a series of 3 experiments, the authors found (a) systematic inferences from small amounts of data; (b) systematic inference of extraneous causal links; (c) influence of data presentation order on inferences; and (d) error reduction through pretraining. Without pretraining, a model based on local computations fitted data better than a Bayesian structural inference model. The data suggest that local computations serve as a heuristic for learning causal structure.

Keywords: causal learning, causal structure, intervention, Bayesian inference, heuristics

Knowledge of causal structure guides explanation, prediction, memory, communication, and control (Sloman, 2005). How is such knowledge obtained? Hume (1777/1975) argued that causal relations are not directly observable and therefore must be inferred on the basis of observable cues. While he suggested several cues such as contiguity and temporal order, it was the regular co-occurrence of cause and effect, *constant conjunction*, he deemed to be the most important and reliable source of information for causal induction. For example, no spatial or temporal cues seem to pertain to inferring a causal relation between the moon and the tides, yet such a relation was inferred historically on the basis of their covariation. We will argue that people do not rely on covariation when learning the structure of causal relations but instead tend to use a variety of cues that allow causal structure to be built up piecemeal from local links.

Associative Strength Learning

The focus on regular co-occurrence has set the stage for contemporary learning theories that have formalized Hume's notion of constant conjunction in a variety of ways. Associationist theories inspired by the animal conditioning literature (e.g., Rescorla & Wagner, 1972), for instance, have represented causality as a predictive relation based on associative strength. As two events are repeatedly paired, the learner comes to expect that one will be accompanied by the other. According to a correlational model, ΔP,

judgments of causality are based on contingency, the increase in probability of an effect given the cause from the base probability of the effect (Jenkins & Ward, 1965). In the limit of infinitely many learning trials, the Rescorla–Wagner learning rule converges to ΔP (under reasonable assumptions, Danks, 2003). This shows that the contingency judgment can be approximated by a simple iterative algorithm that changes associative strength on each trial, and there is a good deal of empirical evidence for a relation between contingency and judgments of causality (Allan & Jenkins, 1980).

Despite these virtues, associative accounts have limitations as descriptions of human learning. They do not provide a means to distinguish causes from effects, nor do they distinguish genuine causal relations from correlations due to common causes. Moreover, people (Hagmayer, Sloman, Lagnado, & Waldmann, 2007) and even rats (Blaisdell, Sawa, Leising, & Waldmann, 2006) can distinguish observing an event from intervening to produce that event when learning causal structure. Intervention has been deemed the hallmark of causality (e.g., Pearl, 2000; Woodward, 2003) in the sense that A causes B if and only if a sufficient intervention to change the state of A would also change B (in the absence of anything to disable the effect of A on B). While an associative model might capture such a distinction post hoc, the current state of the art deploys different associative frameworks to represent observational (Rescorla & Wagner, 1972) and instrumental (Colwill & Rescorla, 1986) learning. Parsimony dictates that capturing both in a single framework would be preferable.

Causal Strength Learning

For these reasons, subsequent theories have proposed causal representations as opposed to associative ones. Cheng's (1997) power PC theory (short for a causal power theory of the probabilistic contrast model) assumes an a priori assignment of causal roles to variables of interest and a causal learning process that aims to induce the *causal power* of the putative cause to bring about the effect. In this theory, causal power is construed probabilistically,

as the extent to which the putative cause changes the probability of the effect in the absence of other causes. This consideration of alternative causes solves many of the problems of associative theories, as the power of a cause to bring about an effect is judged by considering only cases in which the effect would not have happened otherwise.

ΔP and power PC and their variants have been successfully applied to model many causal learning experiments although there is an ongoing and vigorous debate concerning which theories have the most empirical support (see Hattori & Oaksford, 2008). One limitation of theories of this type is their lack of generality. Causal learning is conceptualized as inducing the strength of the link between a single target cause and effect, with the causal role of the cause and effect specified a priori. However, causal learning often requires inducing causal knowledge beyond the strength of a particular relation; a learner might want to induce the causal structure relating several variables rather than or in concert with causal strength.

Causal Structure Learning

Whereas a causal strength inference aims to induce the strength of a particular causal relation that is already known, a structural inference induces how multiple variables causally relate: what causes what, and how causes combine to bring about their effects. This kind of information can be represented in a *causal Bayesian network* (Pearl, 2000), a graphical representation of the causal structure of a domain that also represents the probability distribution that describes events in the domain. More formally, causal Bayes nets are graphs composed of nodes representing events or properties and directed edges representing causal relations between those events or properties, with the network parameterized to represent the joint probability distribution defined over all values of the nodes. Because causal Bayes nets are consistent with probability theory, inferring their structure and parameters can be accomplished by statistical algorithms that attend to the patterns of covariation among the events and properties. Various structure-learning algorithms have been proposed as psychological theories of causal learning (Gopnik et al., 2004; Griffiths & Tenenbaum, 2005).

The ΔP and power PC equations can be interpreted as parameter estimations over causal Bayes nets (Glymour, 2001; Griffiths & Tenenbaum, 2005). For example, consider a causal structure with a target cause, an effect and an alternative cause (a common effect model), in which the causes are related to the effect by a *noisy-or* gate. A noisy-or is a probabilistic version of an inclusive-or. Each cause alone leads to the effect with some probability and when multiple causes are present, the likelihood of the effect is even higher, increasing according to the independent contribution of each cause. In such a situation, the power PC equation is the maximum likelihood estimate of the causal strength of the link from target cause to effect. The noisy-or gate or the closely related multiple sufficient causes schema (Kelley, 1972) is central to many types of causal inferences. This could help explain causal power's empirical support.

Irrespective of the particular learning algorithm, the notion that causal learning entails inferring the structure and parameterization of a causal network on the basis of covariation is a powerful idea. It is general in the sense that inferences about causal structure and strength are addressed naturally within the same framework, and the framework can be accommodated to any background assumptions about functional relations between events such as generative, preventive, and enabling causes. The approach is also flexible in that it allows for learning complex causal structures absent specific a priori knowledge. Further, predictions in this framework are consistent with probability theory. This gives the framework motivation on the assumption that people's causal beliefs about events are to some extent adapted to the probabilities of those events.

Problem of Complexity

These approaches all agree that covariation is the primary input to the learning process, but they are unlikely to offer plausible models of human learning if calculating covariation is computationally challenging. Learning the strength of relation from a single putative cause to a given effect can be accomplished with relatively simple learning algorithms, though even in such cases people's judgments are quite variable, especially when data are presented serially (e.g., Buehner, Cheng, & Clifford, 2003). As the problem is generalized to learning causal structure and strength with more than two variables, the computations required become psychologically implausible and even intractable. For example, learning the structure of a two-variable system necessitates considering three possible acyclic structures. Adding just one more node increases that space of possibilities to 25. Causal systems in the world have many more variables than that. Consider learning how a bicycle works or how to play a video game.

The problem of complexity is often underappreciated because learning models are usually tested in contingency learning paradigms in which the task is to determine the strength of a single causal link between a prespecified cause and effect. However, several recent experiments have tested people's ability to make somewhat more complex causal structure inferences from observations of covariation. These experiments have varied in whether causal roles are predefined, in the number of variables in the system, in the mode of data presentation, and in the cover stories. Yet a common finding is that given just observational data, participants' inferences are strikingly suboptimal.

In Lagnado and Sloman's (2004) Experiment 1, participants were asked to infer the structure of a three-variable probabilistic system by observing trials of covarying events. Participants observed the system 50 times, sufficient in principle to recover the true structure, yet only 5 of 36 participants chose the correct model, a proportion consistent with chance responding. Steyvers, Tenenbaum, Wagenmakers, and Blum (2003) tested causal structure inferences using a similar set-up. In the observation phase of Experiment 2, participants chose the correct structure only 18% of the time. Because selections were scored as incorrect even if the selected model fell into the same Markov equivalence class as the correct model and was therefore indistinguishable on the basis of the observational data, this 18% was significantly higher than chance and better than Lagnado and Sloman's participants but still well below optimal. A third example comes from White's (2006) Experiment 1. Given the suboptimal performance in these other experiments, he used a simplified paradigm with fully deterministic causal links and observational data in the form of written sentences expressing co-occurrence of population changes for dif-

ferent species in a nature reserve. Again, performance was no better than chance.

Evidently, observation of covariation is insufficient for most participants to recover causal structure. This casts doubt on purely covariation-based accounts as comprehensive models of causal learning. Power PC theory and learning algorithms for causal Bayes nets are usually construed as computational-level accounts (Marr, 1980), intended to express an optimal solution to the computational problem the system is trying to solve, and they allow for psychological processes that only approximate these optimal computations. But computational models only have psychological plausibility to the extent that they can account for behavior, and if people really are deficient at computing causal structure from covariation, then these theories are at best incomplete.

Causal Learning Is Local

We propose that causal structure learning is accomplished by *local computations*. The locality of causal learning has two related aspects. First, causal learning is *structurally local*. When faced with a complex learning problem involving several variables, people break the problem up by focusing on evidence for individual causal relations rather than evidence for fully specified causal structures. Structural inferences are accomplished by combining these local inferences piecemeal. Second, learning is *temporally local* in that people tend to prefer cues to structures that are easily accessible and do not tax computational resources like short-term memory and attention. Thus cues to structure that are available at a particular point in time are preferred over those that involve aggregating information over many trials. Causal structure is inferred from a series of observations by serially updating a single hypothesized causal model that is composed of the union of all the locally learned causal relations.

This approach can account, first, for the fact that people learn causal relations from small amounts of data. Often, a single trial is sufficient. Second, like many domains of human cognition, causal learning is subject to systematic and counternormative biases. Finally, people use cues beyond covariation, like temporal information, mechanistic knowledge, and interventions to learn causal structure. In line with these findings, local computations allow for rapid learning because of their local focus, they need not respect normative prescriptions, and they are sensitive to cues other than covariation, cues available on individual learning trials. Local computations make minimal demands on memory, as only a single hypothesis needs to be maintained, namely the structure composed of all the locally-learned connections, and statistical information need not be tracked or aggregated over trials.

Waldmann, Cheng, Hagmayer, & Blaisdell (2008) propose a single-effect learning model to explain two findings that provide compelling evidence for structural locality. First, rats exhibit a particular type of second-order conditioning (Pavlov, 1927/1960). When a unitary stimulus is paired with two different effects on separate trials in training, a rat will learn a dependence between the two effects despite the fact that they are presented separately. An example is provided by Blaisdell et al. (2006) who trained rats with interleaved trials of a light followed by food and a light followed by a tone. The tone and food never appeared together on any training trials so they were strongly anticorrelated. Yet, in a test trial when the tone was presented by itself, the rats looked for food. One interpretation of this finding is that the rats reasoned diagnostically to the presence of the light from the tone and then causally to the presence of food from light. In other words, rats appeared to make inferences consistent with a common-cause model despite the strong anticorrelation of the effects in the training. According to the single-effect learning model, this result can be explained by assuming that the rat learns each cause–effect relation one at a time in training and that causal inferences in test trials are accomplished by chaining together inferences about individual relations. This local learning strategy may be due to processing limitations. In early trials, attention to the presence of the cause and the effect requires significant resources and rats fail to consider the absence of the other effect. After many trials, however, rats do show conditioned inhibition between the effects implying that resources are freed up to attend to the anticorrelation (Yin, Barnett, & Miller, 1994).

Further evidence comes from human learning. Hagmayer and Waldmann (2000) trained participants on data from a common cause whereby a genetic mutation produced two substances. As in Blaisdell et al.'s (2006) experiments, participants learned about the links one at a time. They were told that the two substances were studied at separate universities and then observed trials in which the mutation was present or absent and one or the other substance was present or absent. After learning, they observed new cases with or without the mutation and had to judge in each case whether the two substances were also present. A judgment of the correlation between the two substances was recovered, and the correlation was positive, suggesting that participants had local knowledge of individual causal relations that they combined to make trial-by-trial predictions. However, when in a second task, participants judged the frequency of the second substance conditioned on a set of trials in which the first substance was either present or absent, they showed no awareness of the correlation. One interpretation is that they did not understand explicitly that the local causal beliefs implied a correlation or how to express this correlation as a frequency judgment.

The local computational framework that we espouse here is more general than Waldmann et al.'s (2008) single-effect learning proposal. First, according to the single-effect learning model, links are learned by computing causal power over a series of training trials. We are not committed to such a computation. In our experiments, participants observed small data sets, insufficient for reliably estimating causal strength, and still made systematic structural inferences. This difference may be a function of the type of tasks that the two theories are aimed at describing. We focused on cases in which participants have some a priori knowledge about causal strength and then are asked to make structural inferences explicitly on the basis of a small number of observations. Waldmann et al. (2008) focused on cases in which participants had some a priori constraints on causal structure, receive a lot of training data, and then are asked to make trial-by-trial predictions. Under those conditions, people may be motivated to deploy computational resources to compute or approximate causal strength. Here we merely claim that doing so is not necessary for recovering simple causal structures. We do not believe that causal strength estimation from large numbers of identical trials is the most important function of "ordinary" causal learning. Systems in the world are often complex, and the data people observe are sparse

and noisy. Under these conditions the motivation of causal learners is to simply determine which causal links exist rather than to estimate strength.

Second, we are not committed to the constraint that only single links can be learned from a given observation. In our experiments, participants sometimes made inferences about multiple links simultaneously, as when two effects of a common cause were simultaneously present. Rather than constraining the computation in terms of the number of links, we did so in terms of the notion of locality, proposing that structural inferences were restricted to the parts of the structure that were active or informative on a given trial. This allows local computations to make predictions about a more general class of structure learning problems.

Third, we focused on cues other than covariation as the input to the learning process. As these cues tend to rely on the close connection between causal and temporal relations that emerges from the fact that effects never precede their causes, we referred to the use of these cues collectively as *temporal locality*. The importance of cues beyond covariation has long been acknowledged in philosophy, for instance by Hume (1777/1975) and increasingly by psychologists (Einhorn & Hogarth, 1986; Lagnado, Waldmann, Hagmayer, & Sloman, 2007; Shultz, 1982). Compared with covariation, cues such as temporal order, mechanistic knowledge, and intervention provide information that is more readily accessible and less computationally taxing. Indeed, when put into conflict with covariation data, people tend to base causal inferences on these other cues, even when they are fallible (Ahn & Kalish, 2000; Kuhn & Dean, 2004; Lagnado & Sloman, 2006; White, 2006). Other cues are also important in the single-effect learning model as they are used to differentiate cause from effect in the learning phase. For example, Blaisdell et al.'s (2006) rats likely used the temporal order of the light, tone, and food in training to induce causal directionality.

Relation to Bayesian Models

Bayesian learning methods provide a counterpoint to *local computations* because they are global in the sense of representing causal structure learning as an optimal inference that compares fully specified causal structures on the basis of covariation data (Anderson, 1990; Griffiths & Tenenbaum, 2005; Steyvers et al, 2003). In a Bayesian inference, belief about which structures are more plausible a priori is combined with evidence for particular structures provided by new data. Assuming an appropriate likelihood function, Bayesian models are prescriptive and thus serve as a useful benchmark for human performance. We see little reason to treat them as descriptively correct (Sloman & Fernbach, 2008).

Unlike Bayesian inference that explicitly represents uncertainty over all hypotheses, a local computation considers only one hypothesis at a time. The hypothesis is constituted by the union of all causal links implied by local cues on individual trials. This distinction is key to distinguishing the two approaches empirically. The lack of a hypothesis comparison process in local computations allows for conditions under which structural inferences are inconsistent with statistical norms, such as extraneous links being inferred when a simpler structure explains the data. The learner may not realize that a simpler hypothesis is also consistent with the data because only one hypothesis is considered. A particular instantiation of this phenomenon is explored in the experiments.

Another difference is that according to local computations, the structure hypothesis is built up over a series of trials. In the absence of relevant prior knowledge, participants will never infer links for which they have no evidence. This may seem like an obvious desideratum of any learning process, but whether one's default assumption is the presence of a causal relation or the absence of one may depend on how common one believes that causal relations are. This kind of a belief can be embodied in the prior distribution of a Bayesian model. In the case of local computations, the learning process itself enforces parsimony.

Overview of Experiments

To test the local computation idea, we offered participants in our experiments a cue that allowed them to reconstruct a generative causal structure by making local inferences from data. We used a cue that has been an active target of causal learning research, intervention (cf., Lagnado & Sloman, 2004; Schulz, Kushnir, & Gopnick, 2007; Steyvers et al., 2003). Participants observed interventions on a system of three slider bars, and the slider that was intervened on was identified. Because sliders could only move if intervened on or if their causes were active, interventions provided implicit temporal information, namely that any active sliders must have moved after the intervened-on slider. We hypothesized that participants would compute locally, using the implicit temporal cue provided by the intervention as a guide to piece together causal structure over multiple observations.

Causal chains offer a good test case for local computations. Chains are networks with a link from one variable to another and from the second to a third, and so on, depending on the length of the chain. Consider a simple three-variable causal chain in which A causes B and B causes C. Interventions on the root variable, A, will tend to activate both B and C, B directly and C indirectly through B. Absent any other cues, the implicit temporal cue is a faulty guide to the relation between A and C, because it implies a direct relation while in reality the relation is indirect. The implicit temporal cue suggests that the root variable A happened first and that B and C were effects, but it provides no information about the relation between B and C. Thus, a learner using local computations would spuriously infer the existence of a link between A and C. Subsequent interventions on B that activated C would lead to inferences of a link from B to C, and given sufficient data, the learner would infer a causal structure including both links of the chain and a link from the root variable to the terminal variable. We refer to networks that have this form as *confound models*. Examples of a chain model and a confound model are shown in Figure 1. A common cause model is also shown for comparison.
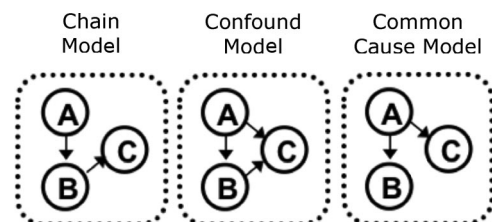

Figure 1. A chain model, a confound model, and a common cause model.

All three experiments investigated induction of causal chains from observations of interventions on a system of causally related slider bars. In Experiment 1, we compared inferences given data from different generative models, common causes, and chains. The implicit temporal cue provides a reliable cue to structure in the former but not the latter case, and we predicted that inferences from common cause data would be close to optimal while participants would systematically infer confound models given data from causal chains. In Experiment 2, we explored whether the order of data presentation could affect inferences (cf. Ahn & Dennis, 2000; Collins & Shanks, 2002; Dennis & Ahn, 2001; López, Shanks, Almaraz, & Fernández, 1998). In accordance with our Bayesian model representing a principled learner who considers all hypotheses, all data were treated equally. Given the serial nature of local computations however, the hypothesis under consideration was different at different points in the series of interventions. This implied that people may make different inferences depending on how data bear on the current hypothesis. In Experiment 3, we explored the conditions under which participants could induce chain models. We varied the instructions and practice trials to push participants to take a more global perspective by explicitly teaching them about the possibility of different causal models.

Experiment 1

In Experiment 1, we compared learning of chains and common cause models. The implicit temporal cues provided by interventions on a common cause are a reliable guide to structure. An intervention on the root variable will tend to activate both of the other variables implying the appropriate causal relations while interventions on the other variables will have no effect. Local computations should result in inferences of common cause models with very small amounts of data.

In the case of chain models, the local computations model leads to inferences of confound models, as discussed earlier. A confound model in this case is not inconsistent with the data. Confound models make qualitatively the same predictions as chain models with respect to interventions on the root variable and the intermediate variable. But the link from the root variable to the terminal variable is not necessary to explain the data. Absent data to choose between the two on quantitative grounds, a principled learner might choose the confound model if he or she believes that causal links are common or might choose the chain if he or she believes they are rare.

To distinguish local computations from a more principled Bayesian computation with a preference for complex structures, we included trials in which the data were insufficient to recover the true structure. Sufficient data mean that there is at least one effective intervention on each of the root and intermediate nodes in the chain. If either or both of those interventions are not observed, the local computations model predicts that participants should infer the structure with the fewest links consistent with the data. Conversely, a Bayesian model with a preference for complexity favors structures with additional links.

For example, consider a single intervention on the intermediate variable that activates the terminal variable. According to local computations this implies a causal structure with a single link from the intermediate variable to the terminal variable. According to a Bayesian model that considers all hypotheses, this intervention is

equally consistent with several models such as the model inferred by local computations, a causal chain, a common cause, and a confound model. The choice between them is dictated by the prior. Learning from local computations thus makes the distinct prediction of confound model inferences given sufficient data from a chain model, along with a general preference for simpler structures in other cases.

To test these predictions, we presented three binary slider bars on a computer screen, and participants were told that their task was to identify the "hidden connections between the sliders that cause some sliders to move when others move." This task was chosen to minimize a priori assumptions about causal structure or mechanisms created by a cover story. Varying the causal model generating the data across trials was simply a matter of changing the conditional probabilities governing slider movements, so we were able to test various causal models with the same materials.

Method

Participants and design. Sixteen Brown University students were paid $5 each for a single session lasting about 1 half-hour. Generative model (chain vs. common cause) was the only independent variable and was tested within participants. Two chain models and two common cause models were used (Figure 2). Each participant completed 40 trials, 10 each for each causal model. We use the word *trial* to refer to a set of five observations of interventions and their effects governed by a particular causal model, and the subsequent causal structure inference. The 40 trials were blocked into groups of four. In each block of four trials, each of the four causal models was used as the generative model once. Presentation order of the four models was randomized in each block.

The dependent measure was which causal model was inferred. Participants were able to choose any combination of causal links that did not produce cycles. They could thus infer one of 25 valid models including one with no links, six one-link models, six chains, three common cause models, three common effect models, and six confound models.

Procedure and stimuli. The experiment consisted of three parts: instructions, practice trials, and experimental trials. Participants first saw brief instructions on a computer screen that told them that they would observe the movements of sliders and be asked to guess the hidden connections between them. They were also informed that connections were probabilistic and directed.

Next they completed two practice trials. We eliminated the possibility of introducing bias by having the practice trials consist of just two sliders. In the first practice trial, the generative model was A causes B and in the other, B causes A. The top half of the screen showed two grey binary sliders, 3 cm wide and 8.5 cm high. Each slider had a 1.5-cm square white label beneath one marked A
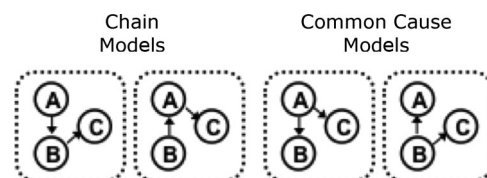

Figure 2. The four generative models tested in Experiment 1.

and the other B. Prior to each intervention, both sliders were in the "down" position. Interventions and slider movements in the practice trials were controlled so that each participant saw the same pattern of data. The label of the intervened-on slider flashed red three times prior to the movement of the sliders and remained red throughout the course of the intervention. In the practice trials, a slight temporal lag was inserted between the movements of the intervened-on slider and other sliders to make the task easier for instructional purposes. The bottom half of the screen showed screenshots of the previous interventions from left to right in the order they occurred. The screenshots were 5-cm square and depicted what appeared during the intervention on the top half of the screen, indicating which variable was intervened on and its effects. Participants pushed a button marked *next* to start the next intervention that began after a 2-s delay.

After observing five interventions, participants were shown a new screen asking them to choose the appropriate arrows indicating the causal relations they ascribed to the model governing the trial. The screen presented A and B labels and two arrows, one in each direction between each label. Participants were instructed to select the appropriate arrows with the mouse. When selected, the arrows turned from gray to red. Feedback provided indicated whether their selection was right or wrong. If the selection was wrong, they repeated the practice trial until they inferred the correct model. After completing both practice trials, they moved on to the experimental trials.

The stimuli and procedure for the experimental trials were identical except that there were three sliders—A, B, and C—during the interventions, there was no lag between the intervened-on and effect variables, and there were three labels and six arrows (between each pair of labels) during the inference portions. Screenshots of the interfaces for observing interventions and inputting inferences during an experimental trial are shown in Figure 3.

Slider movements in the experimental trials were stochastic. Prior to each intervention, the system chose a slider to intervene on at random with uniform probability. The current causal model determined the movements of any other sliders, with the probability of a slider moving given that its direct cause moved set to 0.8. A slider could not move unless it was intervened on or its direct cause was active. No feedback was provided and incorrect trials were not repeated.

At all times there was a trial counter on the top left of the screen indicating how many trials out of 40 had been completed and a button marked *instructions* that the participant could click to view the instructions again.

Results

Model selection. Model selections are depicted in Figure 4. There was no difference between the two chains, $\chi^2(4, N = 16) = 6.3$, $p = .18$, or common causes, $\chi^2(4, N = 16) = 1.6$, $p = .81$, so we collapsed the data for each pair. When the generative model was a chain, confound models and zero- and one-link models were predominant. When generative model was a common cause, common cause models and zero- and one-link models were predominant. The distributions in Figure 4 were compared to a chance model defined as uniform selection from the set of possible models. Chi-square goodness of fit tests revealed significant deviations of chain responses, $\chi^2(4, N = 16) = 86.4$, $p = 0$, and common
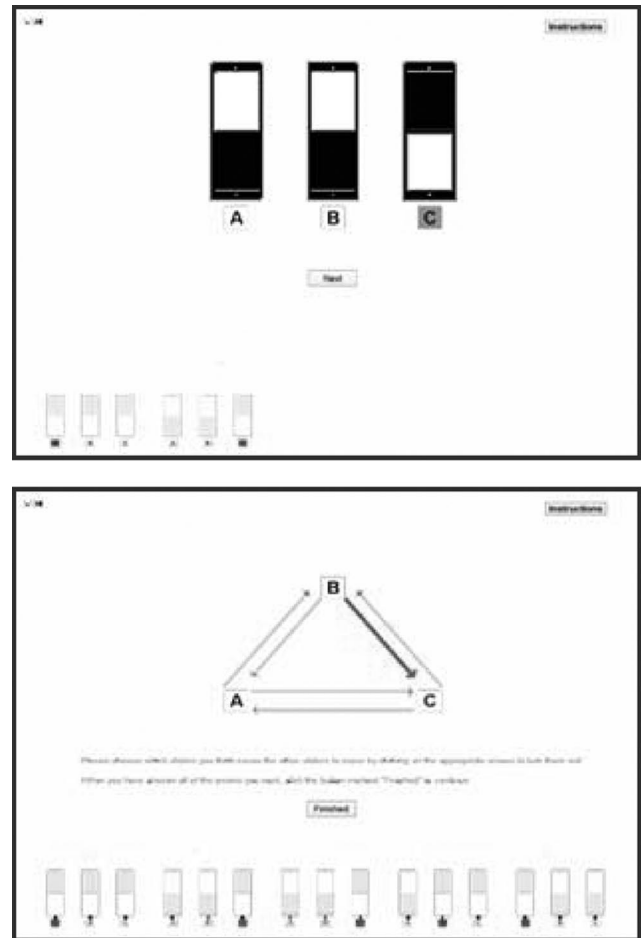

Figure 3. Top panel: Screenshot of the interface for observing interventions, shown after the second of five interventions. Bottom panel: Screenshot of the interface for inputting inferences, shown after the participant has chosen a single link from B to C.

cause responses, $\chi^2(4, N = 16) = 715.9$, $p = 0$, from chance, and from each other, $\chi^2(4, N = 16) = 209.0$, $p = 0$.

The learning data in this experiment were randomly selected and therefore varied across participants. To get a quantitative measure of performance that takes the data observed into account, we compared learning models. We fit responses to a simple heuristic model on the basis of local computations and to a Bayesian model optimizing a particular objective function.

Heuristic model. According to the heuristic model based on local computations, the intervention is treated as an implicit temporal cue and a direct causal relation is asserted between an intervened on variable and any other active variables. Causal structure is built up over the course of the five interventions serially by combining the links inferred on each intervention.

According to the heuristic model, participants should have difficulty inferring chains but should have no trouble inferring common causes given sufficient data. For example, in the case of an A-causing-B-causing-C chain, a participant would rarely get evidence for A-to-B and B-to-C links without simultaneous evidence for an A-to-C link because A and B would rarely both be active in
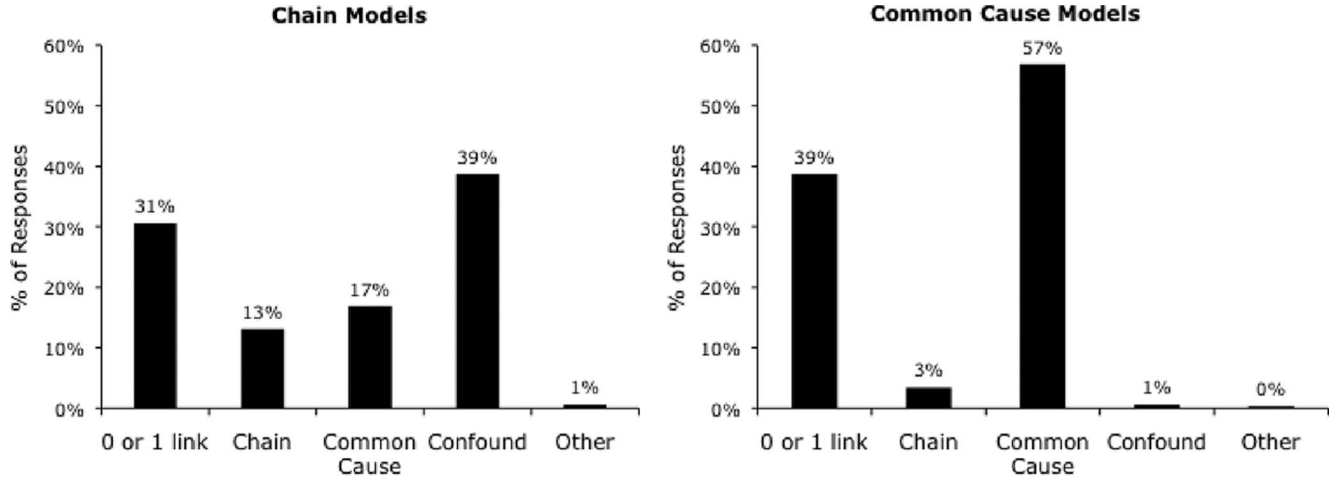
Figure 4. Model selection results for trials on which the generative model was a chain and for trials on which it was a common cause.

the absence of C. Thus, participants would tend to infer confound models that included both links in the chain plus a link from the root to the terminal variable. In the case of common cause models, the participant would tend to see data consistent with a common cause model and would never observe data consistent with spurious links. If the root variable was intervened on, the participant would observe data that implied a link from the root to the other variables, as in a common cause. If the other variables were intervened on, no additional variables would be activated. For both common cause and chain models, the heuristic model predicted that given insufficient data to recover the true two-link structure, participants would tend to infer the zero- or one-link structure that is consistent with the data.

Bayesian model. The Bayesian model represents a principled statistical inference of the best structure given the data and a priori beliefs about which structures are more probable. By Bayes' rule, the posterior probability of a hypothesis, h_i, given data, d, is as follows:

$$P(h_i|d) = \frac{P(d|h_i)P(h_i)}{\sum\limits_{h \in H} P(d|h)P(h)}, \tag{1}$$

where $P(d|h_i)$ is the likelihood of the data under the hypothesis, that is, how probable the given data are under the given hypothesis. $P(h_i)$ is the prior probability of the hypothesis, the degree of belief in the hypothesis before any data has been seen. The denominator normalizes across all hypotheses, H, to generate a probability distribution.

The model is initialized with a hypothesis space consisting of all 25 possible acyclic, three-node causal models. Because inferences are made after only five interventions, different models will often have the same likelihoods. Choosing between these models requires an inductive bias. Consider a single intervention on B that activates C. This outcome is equally likely under a number of hypotheses, such as a one-link model with a link from B to C and a causal chain from A to B to C. There is simply no information about the effects of A because it has not been active. In such a case, it is reasonable to choose

between these models on the basis of how likely one thinks that causal links are in general. This belief can be represented in the model as a parameter that determines the prior probability of a hypothesis on the basis of how many links it has; that is, an a priori bias for simpler or more complex structures. This parameter was fit to the data and the prior distribution was calculated according to the following equation:

$$p(h_i) = \frac{\theta^{-l_i}}{\sum\limits_{i=1}^{25} \theta^{-l_i}}, \tag{2}$$

where $p(h_i)$ is the prior probability of the ith model, l_i is the number of links of the ith model and θ is the parameter representing a bias to simpler or more complex models. The denominator is a constant that normalizes the prior to a probability distribution. A value of θ greater than 1 results in a distribution with higher prior probability on structures with few links (a simplicity bias) while a value of θ between 0 and 1 favors structures with more links. A value of θ equal to 1 results in a uniform prior.

We calculated the likelihoods using the true parameterization from the experiment in which the probability of an effect in the presence of its cause was set to 0.8 and the probability of an effect in the absence of its cause was 0. Parameterizing the model with particular values of these conditional probabilities drastically simplifies the inference problem because parameter values do not have to be estimated simultaneously with structure (cf. Steyvers et al., 2003). In the case of a common effect structure, incoming links were treated as combining via a noisy-or function.

The model represents an experimental trial consisting of an inference after five interventions. The model calculates the posterior distribution over hypotheses after the first intervention using Bayes' rule and the prior distribution and likelihood, and Bayesian updating continues over subsequent interventions with the posterior distribution from the previous intervention as the prior. This is equivalent to calculating the posterior by multiplying the prior by the joint likelihood of all the data rather than updating serially because the interventions are independent. Thus the posterior

probability of each structure can be calculated according to Equation 3 as follows:

$$P(h_i|j_1, j_2, \ldots j_5) = \frac{P(h_i)\prod_{j \in J} P(j|h_i)}{\sum_{h \in H}\left[P(h)\prod_{j \in J} P(j|h) \right]}, \quad (3)$$

where h_i is the ith structure and J is the set of five interventions.

After all five interventions, the posterior distribution represents the relative degree of belief over causal structures. In order to generate predictions from the posterior, we tested two inference rules, sampling and maximizing. In the case of sampling, the model outputs a sample from the posterior distribution. In the case of maximizing, the structure with the maximum posterior probability is output. In the rare case that more than one structure has the maximum posterior probability, the model samples from those structures uniformly.

Simulation details. To fit the heuristic model, we generated predictions for each of the five-intervention trials observed by participants, and the model was scored as correct if it output the exact structure inferred by the participant. Because the heuristic always outputs the same response given the same inputs and has no parameters, the model only had to be run once for each participant.

Since the Bayesian model samples from the posterior, both in the sampling version of the model and in the maximization version when there are multiple maxima, the number of responses correctly predicted varies from run to run. We therefore ran each simulation 100 times across all the data, ignoring which participant they came from, and took the mean number of responses predicted to represent model performance. Again, a response was scored correct if it matched the participant inference. There was little variance across runs of the models so the mean is a reliable indicator of model performance.

Simulation results. Figure 5 depicts the percentage of responses predicted by each of the models, the heuristic model, the Bayesian model with maximization, and the Bayesian model with sampling. For the Bayesian model, we tested various values of the parameter θ. In general, the best fits had parameter values greater than 1, indicating a preference for simplicity. Varying the parameter at values greater than 1 only led to marginal differences in model performance. Figure 5 shows the results using the best fitting parameters, θ = 9 for the maximization model and θ = 10 for the sampling model.

The best overall fits were obtained with the heuristic model, consistent with the model selection results described earlier. A Z test comparing proportions revealed a significant difference between the proportion of responses predicted by the heuristic model compared with the Bayesian maximization model ($Z = 8.9, p = 0$) and to the Bayesian sampling model ($Z = 10.6, p = 0$). A parameter favoring simplicity in the Bayesian model trades off an ability to account for zero- and one-link responses while failing to predict confound models. Thus, the heuristic's unique prediction of confound models along with a general preference for the simplest models consistent with the data leads to better fits. The Bayesian model obtained better fits with maximization than with sampling at every parameter value greater than .2 due to the fact that participants were relatively consistent in their responses. Using the best fitting parameters, predictions obtained with maximization were better than those obtained with sampling, and this difference was marginally significant ($Z = 1.7, p = .08$).

The models were also fitted to individuals using the overall best fitting parameters, and those fits are shown in Figure 6. For simplicity the results of the sampling Bayesian model are not shown as they were lower than maximization. The heuristic model predicts a high percentage of most participants' responses. Of the 16 participants, 11 were fit better by the heuristic model, and all of those differences were significant. Five were fit better by the Bayesian model because those participants sometimes inferred chain models rather than confound models. However, none of those differences were significant.
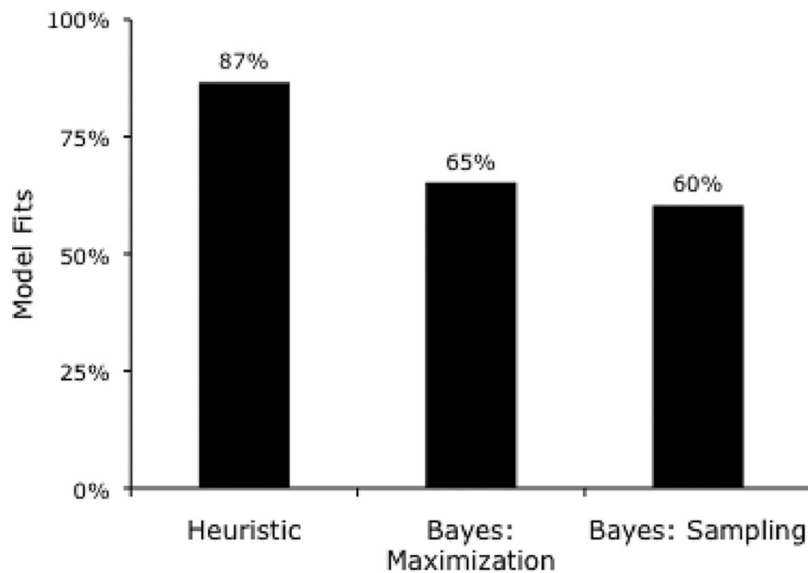

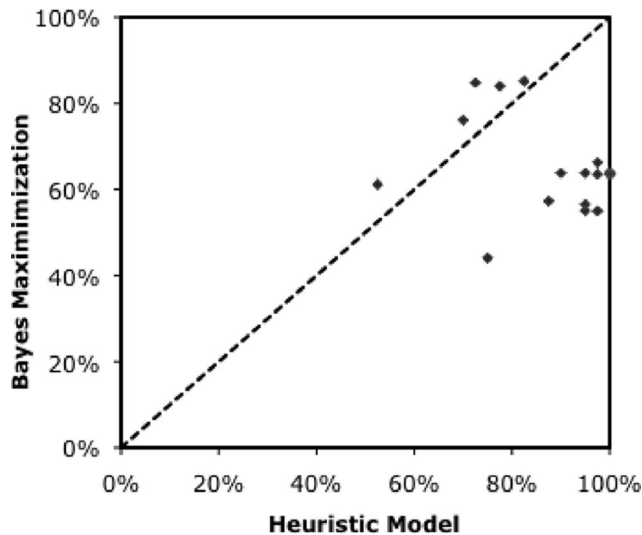*Figure 5.* Percentage of overall responses predicted by each of the models.

Figure 6.   Percentage of responses predicted by the heuristic model and the Bayesian maximization model for each participant.

Discussion

In Experiment 1, we tested people's causal structure inferences given observations of interventions on chain and common cause systems. The results provide strong support for local computations. In the case of common cause models where the implicit temporal cue provided by the intervention is a reliable guide to structure, participants' inferences were close to optimal. In the case of chains, participants systematically inferred an extraneous link from the root variable to its indirect effect, the terminal variable, given sufficient data to recover the true structure. Given insufficient data, participants generally favored the most parsimonious structure consistent with their observations, as indicated by the high correspondence with the heuristic model. The overall pattern of results was predicted by the heuristic model, which uses the implicit temporal cue from each observation as evidence for a local causal relation and builds up causal structure piecemeal over the course of observations. It was inconsistent with the predictions of a Bayesian model representing an optimal statistical learner.

One striking aspect of the results is the degree of systematicity in responses despite only five observations supporting each inference, a pattern contrary to many of the causal structure learning experiments discussed earlier. This speaks to the informativeness of the implicit temporal cue provided by the intervention. Evidently, learners find a cue that provides some information on each learning trial very useful.

Participants were generally consistent in that they never inferred chains when they preferred confound models, suggesting that they may not have been aware of the possibility of chains. To test this, we performed a Bayesian simulation with the prior on chain models set to 0 and with a maximization inference rule. The best fits obtained with this model were precisely the same as the fits obtained with the heuristic. Assuming a relatively strong bias for simplicity, such a model makes the same predictions as the heuristic because without any posterior probability on chains, the confound model is the most likely structure given data from a generative chain.

The fact that the local computation heuristic makes the same predictions as a Bayesian computation with a prior of 0 on chain models shows that a simple heuristic can yield responses that are close to optimal when the cues that the heuristic uses are reliable. It also suggests two ways to interpret our findings. On one hand, the heuristic account implies that participants' difficulty in inferring chains stemmed from a process that computed locally. On the other hand, the Bayesian model suggests a statistically optimal learner with an a priori bias against chain models. Computationally, the two ways of understanding the results are isomorphic in the sense that they yield the same predictions. However, the first interpretation explains the findings in terms of the fundamental assumptions of the theory, (i.e., the locality of computation), while the Bayesian model can only account for the results in an ad hoc manner by positing an extra parameter for the prior on chains.

Experiment 2

In Experiment 1, we compared inferences from causal chains with common cause models in order to distinguish local computations from covariation-based models of learning. The design capitalized on the emphasis of local computations on trial-by-trial cues that can be fallible when they suggest spurious causal links. In Experiment 2, we tested a different implication of local computations, namely that people consider only a single causal structure hypothesis. In the global Bayesian computation, the learner maintains a distribution of belief over all hypotheses, updated as new data come in. Because interventions are independent of one another, the posterior distribution after all data have been observed is insensitive to presentation order. According to local computations, only one causal structure hypothesis is maintained, the union of all the locally learned connections, and that hypothesis changes over the course of observations. We hypothesized that inferences may depend on the order of data presentation.

Consider once again the A-B-C chain model. Participants in Experiment 1 tended to infer confound models after observing interventions on this model. The confound model is not the most parsimonious explanation, but it does account for the data. Participants were unlikely to see any data that would allow them to distinguish between chains and confounds over the course of five interventions because an intervention on A tended to activate both B and C. Evidence for an A-to-B link was almost always concurrent with evidence for an A-to-C link. An intervention on A that activated just B would have been very unlikely, occurring on only 16% of interventions on A, but it also would have been highly diagnostic of chains. Given a confound model, the likelihood of seeing an intervention on A that activated B but not C was much lower, because two causes must fail to bring about their effects.

Consider the presentation orders of interventions in Table 1. In the first case, the diagnostic intervention occurs at the beginning of the series. According to local computations, the learner has not made any commitments about causal relations in the model. The diagnostic intervention is still diagnostic of chains over confound models in that it supports a link between the root and the intermediate variable in the absence of a link between the intermediate and terminal variables, but it is not inconsistent with the learner's current beliefs. In the second case, however, the diagnostic intervention happens at the end when the learner has already built up a

Table 1

Presentation Orders of Interventions on Causal Chains in Experiment 2

	Presentation order 1		Presentation order 2	
Intervention	Active variables	Intervention	Active variables	
Root[a]	R, I	Root	R, I, T	
Intermediate	I, T	Intermediate	I, T	
Terminal	T	Terminal	T	
Root	R, I, T	Root[a]	R, I	
Terminal	T	Terminal	T	

Note. R = root; I = intermediate; T = terminal. In the first condition, the rare but diagnostic intervention in which the root and intermediate variables are active in the absence of the terminal variable occurs early. In the second condition, it occurs late.

[a] Diagnostic intervention.

confound model and has beliefs that are inconsistent with a confound model. This is the difference we tested in Experiment 2.

There are four likely patterns of responses each consistent with a hypothesis about how a learner might use the diagnostic intervention. First, a principled learner calculating the likelihood of the data under various hypotheses would infer chains regardless of when in the series the diagnostic intervention occurs assuming no strong a priori bias for confound models. Second, a pure-heuristic learner, using the implicit temporal cue as the sole guide to structure and without considering how new data bear on the current hypothesis, would infer confound models regardless of presentation order. In both of these cases, no order effect is predicted.

The third case is a recency effect, in which late presentation of the diagnostic intervention increases chain inferences. This pattern is consistent with what we refer to as an *accumulator*. A learner of this type uses local computations to determine a causal structure hypothesis after each intervention and then engages in an assessment process whereby the current data are compared to the current hypothesis. The hypothesis may then be revised if the data are sufficiently surprising or inconsistent with the current hypothesis. In the case of the late presentation of the diagnostic intervention, early local computations would lead to a hypothesis of a confound model. The diagnostic intervention may cause the learner to reassess the hypothesis and revise it to a chain model achieving greater consistency with the data. In the case of early presentation, the diagnostic intervention is consistent with the hypothesis derived from local computations, a model with one link from the root to intermediate variable, and no revision is necessary.

The fourth case is a primacy effect where early presentation increases chain inferences. This effect is consistent with a "hypothesis tester." Such a learner is aware of which variables could be active given his or her hypothesis. On a given intervention the learner looks for active variables that are not predicted by the model and adds links according to local computations to account for any unpredicted variables. In such a case early presentation of an intervention on A that activates B and an intervention on B that activates C leads to a hypothesized chain structure via local computations. The subsequent intervention on A that activates all the variables is predicted by the chain hypothesis and no links are

added, resulting in a chain inference. When the diagnostic intervention is presented late, the extraneous link from A to C is asserted after the first intervention, when all variables are active. The accumulator is liberal in asserting links but can then remove them if necessary, while the hypothesis tester is conservative in asserting links in the first place.

Method

Participants and design. Twenty-five Brown University students were paid $5 each for 1 half-hour session. Participants were randomly assigned to conditions, 12 to the early group and 13 to the late group. The main independent variable, early versus late presentation of the diagnostic intervention, was tested between participants. The only difference between conditions was the flipping of the first and fourth interventions on chain trials. There was also a within-participants variable, chain versus common cause model. Common cause models were included to create variance in the data observed by participants. Three chain models and three common cause models were tested (see Figure 7). For each group, the pattern of data presented for each chain model was identical. Data from common cause models were stochastic as in Experiment 1. Each participant saw 60 trials, 10 of each causal model. The 60 trials were blocked into groups of six, with each generative model tested once per block. The order of the six models was randomly determined in each block.

Stimuli and procedure. Stimuli were identical to Experiment 1 except that in the chain model trials, the interventions and slider movements were not stochastic but were predetermined as shown in Table 1. The procedure was identical to Experiment 1 except that there were 60 trials instead of 40.

Results

Model selection. Two participants in the late group responded at random and were not included in subsequent analyses. Model selection results for the early and late groups are shown in Figure 8. We compared responses to the three chains and common causes by conducting chi-square independence tests for the distribution of early-group chain responses, $\chi^2(4, N = 25) = 0.3$, $p = .98$; early-group common cause responses, $\chi^2(4, N = 25) = 2.0$, $p = .35$; late-group chain responses, $\chi^2(4, N = 25) = 1.7$, $p = .78$; and late-group common cause responses, $\chi^2(4, N = 25) = 3.1$, $p = .21$. Since there were no differences between models of the same type, we collapsed across chains and common causes as in Experiment 1.

We compared the distribution of responses to chain trials across groups using a chi-square test of independence yielding a significant difference, $\chi^2(4, N = 25) = 106.7$, $p = 0$. Participants in the late group were more likely to infer chains. There was no difference between early and late groups on common cause trials, $\chi^2(4, N = 25) = 5.8$, $p = .21$. The predominance of confound models
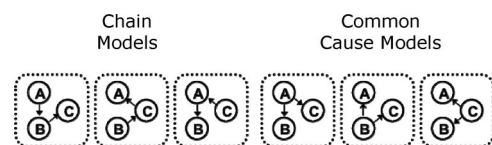

Figure 7. Six generative models tested in Experiment 2.

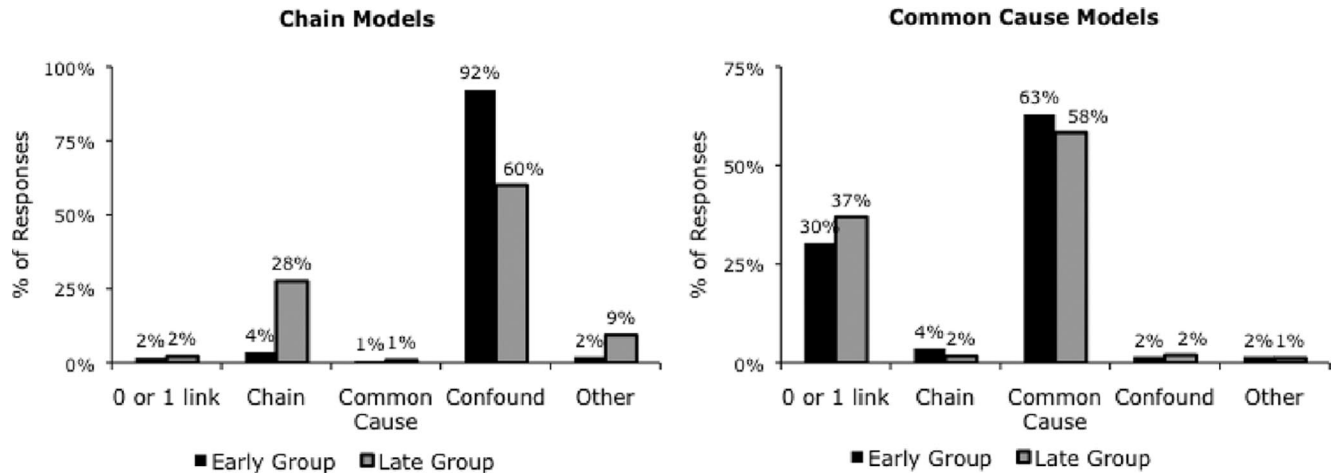**Chain Models**

Common Cause Models

Figure 8. Model selection results for trials on which the generative model was a chain and for trials on which it was a common cause for early and late presentation of the diagnostic intervention.

from Experiment 1 was replicated, with the confound models constituting the majority of responses to chain trials across both groups. To see whether there was a learning effect over the course of the experiment, we tested whether chains were more readily inferred later in the experiment than earlier. A binomial test revealed no such effect. Of the 106 times that chains were inferred, 57 of those were in the first half of the experiment and 49 were in the latter half ($Z = .77$, $p = .44$).

Simulation results. The results were fit to the two learning models as in Experiment 1 (Figure 9). For the Bayesian model, only maximization is shown as it obtained better fits than sampling. The best fitting parameter discovered with the Bayesian model was $\theta = 1$. Note that unlike Experiment 1, the best fitting parameter did not favor parsimony. This happened because on

each chain trial, sufficient data were presented to recover the true structure so on average, participants inferred more confound models and fewer zero- and one-link models than in Experiment 1. This made it difficult for the Bayesian model to obtain reasonable fits with a prior that favored simpler models. Overall, the heuristic fit the data for both the early and late groups better than the Bayesian model and both differences were significant (early group: $Z = 5.9$, $p = 0$; late group: $Z = 4.9$, $p = 0$). Model fits to individual participants are shown in Figure 10.

Discussion

In Experiment 2, we found that the order of data presentation affected causal structure learning. Specifically, participants were
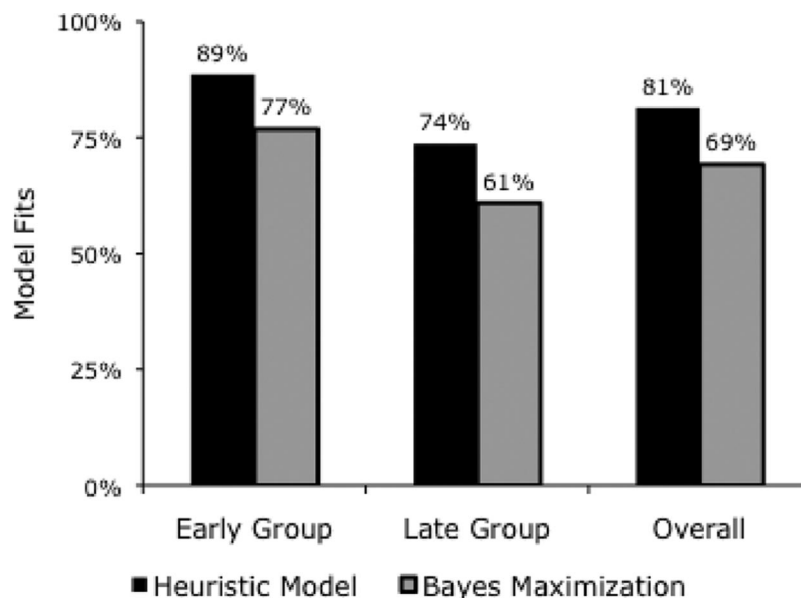

Figure 9. Percentage of responses predicted by each of the models for the early group, late group, and overall.
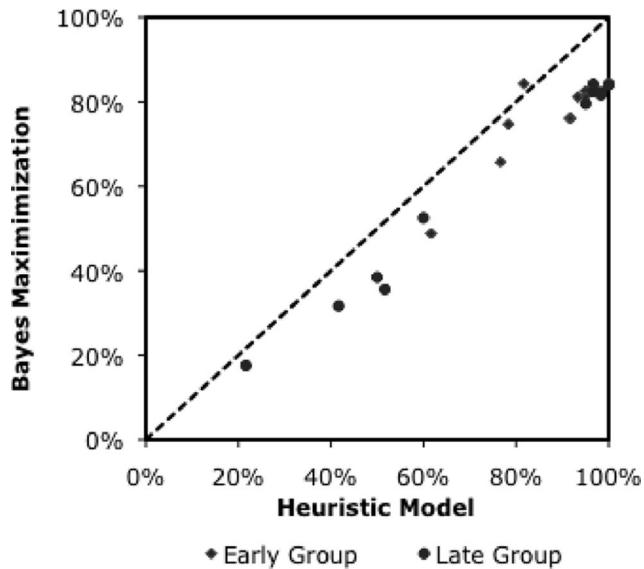
Figure 10. Percentage of responses predicted by the heuristic model and the Bayesian model for each participant, by group.

more likely to infer chain models over confound models when a particularly diagnostic intervention occurred late in learning. We also replicated the predominance of confound model inferences from Experiment 1. Even given the diagnostic intervention, which is highly unlikely to come from a confound model, participants still tended to infer confound models over chains across both conditions. This underscores the powerful influence of local cues on learning and is evidence for local computing. The superior fits of the heuristic model in both conditions also supported this conclusion.

The recency effect that was observed is consistent with an accumulator learner because his or her learning process involves accumulating causal links according to local cues and then revising the hypothesis in the face of inconsistent or surprising observations. The effect of presentation order is not predicted by the simple heuristic described earlier. Rather, the effect can be understood to be the result of an error correction mechanism that reacts to data that disconfirm the locally learned hypothesis, and the differential effect of early versus late presentation of the diagnostic intervention supports the notion that a single hypothesis is updated over the course of learning. The fact that we observed a recency rather than a primacy effect speaks to the automatic application of local heuristics. Our results suggest that causal links are learned liberally via local computations and sometimes are subsequently pruned in the face of inconsistent observations. If participants had used a more active hypothesis-testing strategy, then early presentation of the diagnostic intervention would have yielded more chain responses because in those cases the chain model would have been a good explanation for the activation of the terminal variable, and there would have been no reason to add the spurious link. Instead, participants who showed the order effect revised their hypothesis reactively in the face of data inconsistent with their hypothesis, the failure of the terminal variable to activate.

Despite the presence of an order effect, the simple heuristic was able to account for the majority of responses. The finding that a minority of participants were able to engage in error correction leading to optimal responding implies that given sufficient effort, learners can sometimes adapt flexibly to the available information, a point supported by Steyvers et al.'s (2003) finding of clusters of participants who were substantially better than others at learning from covariation data. The fact that we observed the recency effect hints that a common, low-effort local computation may underlie human competence in general but that some participants are able to deploy a more deliberative strategy to augment performance in the face of counterevidence.

One might argue that a covariation-based model could explain an order effect if it were augmented with a constraint on which interventions enter into the likelihood calculation. For instance one might assume that people attend differentially to late trials and base the calculation of the likelihood on the last n observations. This could explain a recency effect because when the diagnostic intervention comes early in learning, it might be ignored or forgotten. In that case, the relative likelihood of the confound model would be higher than if the diagnostic intervention had come later and had figured in the likelihood computation. We are skeptical about this interpretation for three reasons. First, there does not seem to be any good motivation for predicting a recency effect rather than a primacy effect. If people are limited in the number of interventions they can use for the computation, then they may attend to early trials rather than recent ones. Second, recall that participants could refer back to screenshots of the previous interventions, so memory demands cannot explain systematic disregard for certain interventions. Third, this type of model does not account for the overall pattern of results from Experiments 1 or 2. If there were systematic disregard for certain interventions, we should have observed more sparseness in the causal links inferred by participants. Instead, responses were generally consistent with the heuristic that takes account of all five interventions.

Experiment 3

In Experiments 1 and 2, we found evidence for local computations in the tendency to infer extraneous links given observations of intervention. However, the presence of an order effect in Experiment 2 implied that local heuristics do not by themselves fully account for causal learning. Rather some participants were able to react to evidence contrary to their beliefs and generate responses more consistent with the data. In Experiment 3, we investigated how entrenched the heuristic was. We hypothesized that people are not stuck using local computations but will consider hypotheses from other sources if they are presented explicitly. We therefore "primed" chain models by explicitly teaching participants about them prior to testing, and we predicted that the manipulation would improve performance relative to Experiments 1 and 2.

Two methods for increasing awareness of chains were tested in a between-subjects design. The *handout* group received a description of causal models prior to the experiment. The description included real-world examples of a common cause and a causal chain that were presented graphically. The *practice* group received no handout. Instead, their practice sessions consisted of two models with three variables, a chain and a common cause, rather than two models with two variables as in previous experiments and the handout group of Experiment 3. Participants in the practice group had to repeat the practice trials until they inferred the correct

structure. Thus, they learned to infer chain models in the experimental context.

Method

Participants. Thirty-three Brown University students were paid $5 for one session lasting one-half hour. Participants were assigned randomly to the two groups, 16 to the handout and 17 to the practice group. As in Experiment 1, generative model (common cause vs. chain) was a within-participants factor. Handout versus practice was varied between participants.

Stimuli and procedure. Prior to beginning the computer portion of the experiment, participants from the handout group received a sheet of paper describing causal models and giving examples of common cause and chain structures. The common cause model was that smoking causes lung cancer and yellow teeth and the chain model was that smoking causes lung cancer and lung cancer causes death. Participants were told to read the handout carefully and were observed to make sure they read it. After reading the handout, the procedure was identical to that of Experiment 1.

Participants in the practice group were not given the handout. They received the same instructions as Experiment 1 and began the practice trials. The practice trials consisted of two three-variable models: a chain, B causes A causes C, followed by a common cause, B causes A and C. The practice trials were the same as experimental trials from Experiment 1 except that there was a slight temporal lag to make the inference easier and the data were not randomly generated. Rather the interventions and activated variables were predetermined and always the same. Participants repeated the practice trials until they inferred the correct structure. After completing the two practice trials, the procedure was identical to Experiment 1.

Results

Model selection. Two participants in the handout group responded at random. Their data were not included in subsequent analyses. As in Experiments 1 and 2, there were no differences between models of the same type—handout-group chains: $\chi^2(4, N = 16) = 6.6$, $p = .15$; handout-group common causes: $\chi^2(4, N = 16) = 2.8$, $p = .24$; practice-group chains: $\chi^2(4, N = 17) = 1.7$, $p = .42$; and practice-group common causes: $\chi^2(4, N = 17) = 1.2$, $p = .55$. We therefore collapsed across the two chain and common cause models. Figure 11 shows the distribution of responses for common cause and chain trials for the practice and handout groups. Experiment 1 results are also shown for comparison. Chi-square goodness-of-fit tests showed that all four distributions were significantly different from chance.

Both types of instruction increased the likelihood of correctly learning chain models, with the practice manipulation having a larger effect than the handout manipulation. The distribution of responses to chain trials was compared across groups using a chi-square test of independence that yielded a significant difference between responses in the practice and handout groups, $\chi^2(4, N = 33) = 34.9$, $p = 0$. Responses to common cause trials were not significantly different across groups, $\chi^2(4, N = 33) = 7.8$, $p = .10$. Comparison of responses on chain trials between the handout group and Experiment 1, $\chi^2(4, N = 16) = 23.5$, $p = 0$, and the practice group and Experiment 1, $\chi^2(4, N = 17) = 102.0$, $p = 0$, yielded significant differences. Common cause responses between the handout group and Experiment 1, $\chi^2(4, N = 16) = 3.9$, $p = .42$, and the practice group and Experiment 1, $\chi^2(4, N = 17) = 6.3$, $p = .18$, were not significantly different.

Simulation results. The heuristic and Bayesian models were fit as in Experiment 1. For the Bayesian model, maximization always led to better fits than sampling, so for simplicity, only maximization results are shown. Figure 12 depicts model fits for the heuristic and for the Bayesian model with maximization for each of the groups and for Experiment 1. As in Experiment 1, varying the parameter as long as it was greater than 1 only made marginal differences to the fits of the Bayesian model, and results with the best fitting parameter are shown, $\theta = 8$ for the handout group and $\theta = 7$ for the practice group. As suggested by the model selection results, the Bayesian model's performance improved from Exper-
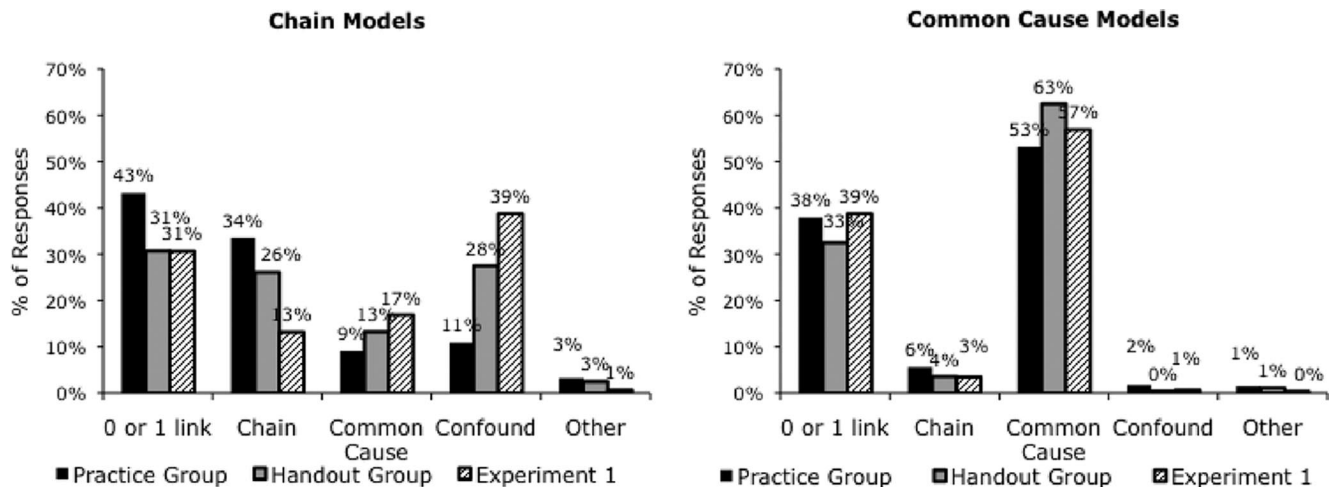

Figure 11. Model selection results for trials on which the generative model was a chain and for trials on which it was a common cause for both Experiment 3 conditions and Experiment 1.
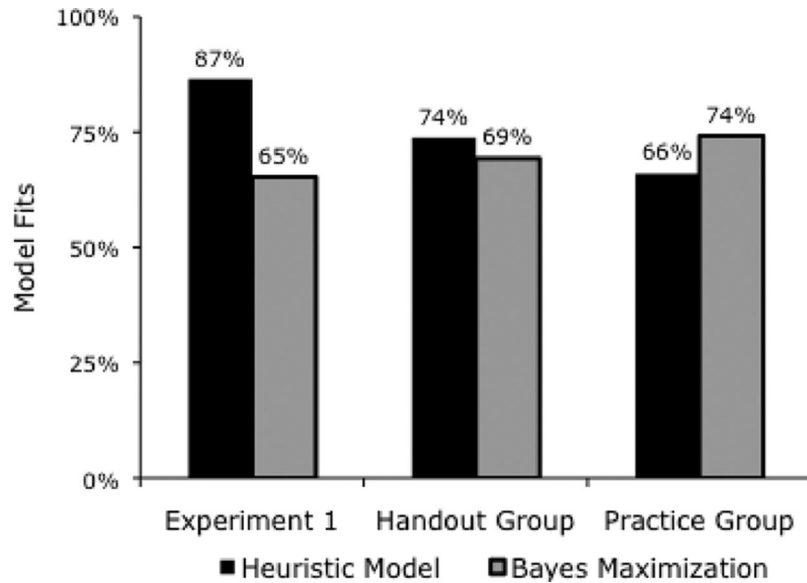
Figure 12. Percentage of overall responses predicted by each of the models for both Experiment 3 conditions and Experiment 1.

iment 1 to the handout group and from the handout group to the practice group due to more chain inferences. For the handout group, the heuristic model predicted the higher overall proportion of responses than the Bayesian model though this difference was not significant ($Z = 1.7$, $p = .10$). For the practice group, the Bayesian model achieved a better fit, and this difference was significant ($Z = 3.1$, $p = 0$).

The models were also fit to individual participants, and those results are shown in Figure 13. In all groups, a subset of partici-

Figure 13. Percentage of responses predicted by the heuristic model and the Bayesian model for each participant in the handout group and the practice group.

pants was fit very closely by the heuristic, implying consistency in inferring confound models over chains; however, the proportion of those participants decreased as chains were primed and chain inferences became more prevalent. Z scores for the proportions of responses predicted were calculated and showed that in the handout group, four participants were fit significantly better by the heuristic and two by the Bayesian model. In the practice group, two were fit significantly better by the heuristic and six by the Bayesian model.
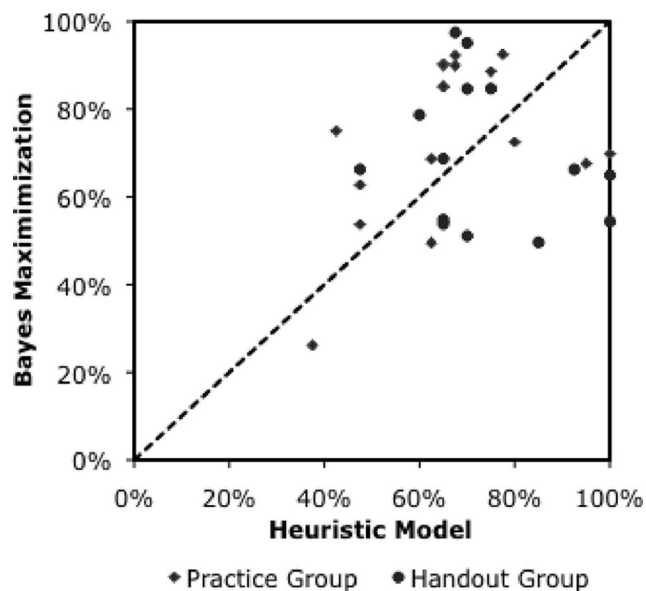
Discussion

As predicted, participants' ability to infer chain models over confound models increased when they were taught about chains in Experiment 3. The effect was stronger in the practice group, implying that learning to infer chains in the context of the experimental materials was more helpful than learning about a particular real-world example. While the overall beneficial effect of training may not be surprising, what is surprising is that despite the priming of chain models, a subset of participants in each group still behaved like the participants in Experiment 1 and was fit well by the local computations heuristics. The results thus reveal when the heuristic model we propose is applicable and when it is not. Local computations are a default that can be overcome by effort or by an environment that provides more support.

The relative success of the Bayesian model when people were taught about chains shows that people were able to learn these simple three-variable causal structures when given a cue to consider one of those structures. The effect of this teaching could be modeled in a Bayesian framework as a change in people's prior probabilities over causal structures, a low prior for chains when they are not taught and a higher prior when they are. Although this is descriptively correct, we do not see that it provides any added explanatory value.

General Discussion

Summary

Over a series of three experiments, we find that people (a) fail to learn chain models because they frequently incorrectly infer extraneous causal links (Experiments 1 and 2); (b) correctly learn common cause models from small amounts of data (Experiments 1 and 2); (c) tend to find support for structural hypotheses that are consistent with the most recent data that they see (Experiment 2); and (d) take into account hypotheses that are explicitly presented (Experiment 3). The data from Experiments 1 and 2 were fit closely by a heuristic model based on the idea that causal structure is built up from trial-by-trial inferences of local relations. Together, the evidence suggests that people use local computations as a simplifying strategy—a heuristic—to learn causal structure from data. Sometimes people use more sophisticated strategies that involve considering structural relations that span more than a single link. Experiment 2 shows that such strategies are more likely in the face of data that are inconsistent with the locally learned structure, and Experiment 3 shows that appropriate training leads to consideration of alternative, simpler causal structures. The current results do not provide unique support for a particular model to explain these deviations from local computations.

Implications

The local computations framework explains why people are poor at using covariation among variables to learn causal structure. When only covariation cues are provided (e.g., Steyvers et al., 2003, Experiment 1), most people lack the resources to make the necessary computations. In contrast, when local cues are provided, they lead to more consistent responding but also sometimes to error. For instance, in Lagnado and Sloman's (2004) experiments, when data were produced by a causal chain (A causes B causes C), people tended to believe that they were produced by a common effect model in which A and B independently caused C. This is consistent with the local computation framework because, due to the cover stories used in those experiments, participants knew which variable represented the ultimate effect in the causal system. Thus, on each trial in which the effect occurred, any other active variable was identified as a cause of that effect. As in our experiments, this inference was faulty, implying independence between the two causes and direct relations between the causes and the ultimate effect.

The local computations framework is also consistent with other causal learning phenomena. For instance, in our experiments, participants tended to infer extraneous causal relations on the basis of local cues. Previous research with pigeons (Skinner, 1948) and humans (Ono, 1987) has shown learning of spurious relations on the basis of temporally contiguous actions and outcomes, even when the outcomes are independent of the actions, a type of learning referred to as *superstitious*. For example, Ono's participants tended to repeat an idiosyncratic series of actions that immediately preceded reward, even when rewards were delivered at a constant rate that was independent of the actions. The local computations account of this finding is that participants used the local temporal cues to learn a relation between the actions and reward while failing to recognize the independence of the two.

Inferring independence would have required tracking and aggregating covariation information over multiple trials.

Another aspect of learning that is problematic for covariation-based accounts but that falls naturally out of local computations is single-trial learning (Guthrie & Horton, 1946). Animals and humans often infer a causal relation on the basis of a single observation of two temporally contiguous events. This is inconsistent with the idea that causal relations are learned by estimating causal strength over many training trials. Rather, causal relations can be learned from even a single observation.

Conclusions

The mix of individual strategies observed in Experiments 1 and 2 and the finding in Experiment 3 that chain hypotheses were more likely to be considered after pretraining on chains shows that local processing can be overcome by consideration of higher order causal structures. This suggests that (at least) two processes are involved in causal learning, a local heuristic process and a more sophisticated one that is able to consider how well global hypotheses fit the data.

Learning causal structure locally is an excellent way to combine multiple sources of knowledge. Some causal relations we learn from doing; others we learn from observation; still others we learn from instruction. Each causal relation can be difficult to learn, especially when it reflects a complex mechanism. Learning relations independently allows us to focus on specific mechanisms while ignoring others, at least temporarily. This may be the only way to learn causal systems involving dozens of variables or more, like social systems, car engines, or word processors. The unfortunate consequence of using a heuristic that minimizes memory and computational demands is that it leads to systematic error. Errors may be exceptional when dealing with an expert, but in this case they very much prove the rule: One can hardly deny the presence of systematic error when people, even experts, are dealing with social systems, car engines, or word processors.

References

Ahn, W., & Dennis, M. (2000). Induction of causal chains. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the twenty-second annual conference of the Cognitive Science Society* (pp. 19–24). Mahwah, NJ: Erlbaum.

Ahn, W., & Kalish, C. (2000). The role of mechanism beliefs in causal reasoning. In R. Wilson, & F. Keil (Eds.), *Explanation and cognition* (pp. 199–226). Cambridge, MA: MIT Press.

Allan, L. G., & Jenkins, H. M. (1980). The judgment of contingency and the nature of the response alternatives. *Canadian Journal of Psychology, 34,* 1–11.

Anderson, J. R. (1990). *The adaptive character of thought,* Mahwah, NJ: Erlbaum.

Blaisdell, A. P., Sawa, K., Leising, K. J., & Waldmann, M. R. (2006, February 17). Causal reasoning in rats. *Science, 311,* 1020–1022.

Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariaton to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1119–1140.

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

Collins, D. J., & Shanks, D. R. (2002). Momentary and integrative response strategies in causal judgment. *Memory & Cognition, 29,* 1138–1147.

Colwill, R. C., & Rescorla, R. A. (1986). Associative structures underlying instrumental learning. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 20, pp. 55–104). New York: Academic Press.

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

Dennis, M. J., & Ahn, W. (2001). Primacy in causal strength judgments. *Memory & Cognition, 29,* 152–164.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin, 99,* 3–19.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology.* Cambridge, MA: MIT Press.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushir, 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,* 354–384.

Guthrie, E. R., & Horton, G. P. (1946). *Cats in a puzzle box.* New York: Rinehart.

Hagmayer, Y., Sloman, S. A., Lagnado, D. A., & Waldmann, M. R. (2007). Causal reasoning through intervention. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 86–100). Oxford, UK: Oxford University Press.

Hagmayer, Y., & Waldmann, M. R. (2000). Simulating causal models: The way to structural sensitivity. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the twenty-second annual conference of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

Hattori, M., & Oaksford, M. (2007). Adaptive non-interventional heuristics for co-variation detection in causal induction: Model comparison and rational analysis. *Cognitive Science, 31,* 765–814.

Hume, D. (1975). *Enquiries concerning human understanding* (3rd ed.). Oxford, UK: Clarendon Press. (Original work published 1777)

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied, 79,* 1–17.

Kelley, H. H. (1972). Causal schemata and the attribution process. In E. E. Jones et al. (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 151–174). Morristown, NJ: General Learning Press.

Kuhn, D., & Dean, D., Jr. (2004). Connecting scientific reasoning and causal inference. *Journal of Cognition and Development, 5,* 261–288.

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

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

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 451–460.

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation: cues to causal structure. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 86–100). Oxford, UK: Oxford University Press.

Marr, D. (1982). *Vision.* San Francisco: W. H. Freeman.

Ono, K. (1987). Superstitious behavior in humans, *Journal of the Experimental analysis of Behavior, 47,* 261–271.

Pavlov, I. P. (1960). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex.* (G. V. Anrep, Trans. & Ed.). London: Oxford University Press. (Original work published 1927)

Pearl, J. (2000). *Causality.* Cambridge, UK: Cambridge University 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 theory and research* (pp. 64–99). New York: Appleton-Century-Crofts.

Schulz, L. E., Kushnir, T., & Gopnik, A. (2007). Learning from doing: Interventions and causal inference. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 86–100). Oxford, UK: Oxford University Press.

Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development, 47,* 1–51.

Skinner, B. F. (1948). Superstition in the pigeon, *Journal of Experimental Psychology, 38,* 168–172.

Sloman, S. A. (2005). *Causal models; how people think about the world and its alternatives.* New York: Oxford University Press.

Sloman, S. A., & Fernbach, P. M. (2008). The value of rational analysis: An assessment of causal reasoning and learning. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 453–484). Oxford, UK: University Press.

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

Waldmann, M. R., Cheng, P. W., Hagmayer, Y., & Blaisdell, A. P. (2008). Causal learning in rats and humans: A minimal rational model. In N. Chater, & M. Oaksford (Eds.), *The probabilistic mind. Prospects for Bayesian cognitive science* (pp. 453–484). Oxford, UK: University Press.

White, P. A. (2006). How well do people infer causal structure from co-occurrence information? *European Journal of Cognitive Psychology, 18,* 454–480.

Woodward, J. (2003). *Making things happen: A theory of causal explanation.* Oxford, UK: Oxford University Press.

Yin, H., Barnet, R. C., & Miller, R. R. (1994). Second-order conditioning and Pavlovian conditioned inhibition: Operational similarities and differences, *Journal of Experimental Psychology: Animal Behavior Processes, 20,* 419–428.