The advantage of timely intervention

Article *in* Journal of Experimental Psychology Learning Memory and Cognition · August 2004

DOI: 10.1037/0278-7393.30.4.856 · Source: PubMed

2 authors:

David Lagnado
University College London
91 PUBLICATIONS **1,417** CITATIONS
SEE PROFILE

Steven A. Sloman
Brown University
141 PUBLICATIONS **6,166** CITATIONS
SEE PROFILE

Journal of Experimental Psychology:
Learning, Memory, and Cognition
2004, Vol. 30, No. 4, 856–876

0278-7393/04/$12.00 DOI: 10.1037/0278-7393.30.4.856

The Advantage of Timely Intervention

David A. Lagnado
University College London

Steven Sloman
Brown University

Can people learn causal structure more effectively through intervention rather than observation? Four studies used a trial-based learning paradigm in which participants obtained probabilistic data about a causal chain through either observation or intervention and then selected the causal model most likely to have generated the data. Experiment 1 demonstrated that interveners made more correct model choices than did observers, and Experiments 2 and 3 ruled out explanations for this advantage in terms of informational differences between the 2 conditions. Experiment 4 tested the hypothesis that the advantage was driven by a temporal signal; interveners may exploit the cue that their interventions are the most likely causes of any subsequent changes. Results supported this temporal cue hypothesis.

Our perspective on the world is inextricably causal. We understand, predict, and control our environment by positing stable causal mechanisms that generate the transitory flux of our sensory experience (Hume, 1748; Mill, 1843/1950; Pearl, 2000; Sloman & Lagnado, 2004). Crucial as such causal knowledge is to our functioning in the world, the details of how we acquire it are unclear. Much is learned through education and instruction and by the use of analogy or boot-strapping from prior causal beliefs. However, some has to be acquired anew, inferred from the changing states of the world and our interactions with it.

Observation or Experiment

Philosophers have long distinguished between two ways in which we learn about causal structure: through observation and experimentation. Mill (1843/1950), when outlining how one discovers causal relations, stated "We may either find an instance in nature suited to our purposes or, by an artificial arrangement of circumstances, make one" (p. 211). More colorfully, in characterizing the role of experiment, Bacon (1620) spoke of torturing nature to reveal its secrets, or of "twisting the lion's tail." The critical difference between the two approaches involves the notion of manipulation. In the paradigm case of observation, one simply registers or experiences patterns of events. In contrast, in the case of experimentation, one actively manipulates some feature of the world and then observes the consequent results. For example, compare learning to use a new software program by watching a preprogrammed video clip with learning to use it through interacting with the program itself.

It is a commonplace intuition that experiment, where possible, is a more effective learning tool than mere observation. Mill (1843/1950) argued that the main way to establish cause–effect relations is by the method of difference, which requires us to introduce or remove a potential cause while keeping other factors constant. He also claimed that experimentation is necessary to identify a unique causal structure. For example, suppose that police statistics reveal that the incidence of drug use and the level of petty crime are highly correlated. Such observational data will typically be insufficient to determine whether drug use promotes petty crime or vice versa. However, if one of these factors is manipulated, that is, by a police initiative to reduce drug usage, a subsequent drop in the crime rate tells us that drug usage is driving the crime rate. More generally, the ability to experiment allows us to discriminate between causal structures that are indistinguishable through observation alone.

Mill's (1843/1950) insights are now implicit in contemporary experimental design. It is common to distinguish between observational and experimental studies and to prefer the latter when possible (Shadish, Cook, & Campbell, 2002). On a grander scale, the claim has been made that a hallmark of modern science was the introduction of the experimental method and hence the passage from uncontrolled observations to controlled experiments (cf. Hacking, 1983).

However, do similar conclusions apply to our everyday learning and reasoning? Intuitively the answer is yes: We are continually conducting informal experiments of our own to learn about the world around us. We remove items from our diet to see what is making us unhealthy or overweight, we tinker with new software programs to see what does what, we experiment with different ingredients in search of a tasty meal. Furthermore, the claimed advantage of intervention over observation resonates with the received wisdom that we learn better if we are able to interact with and explore our learning environment rather than just observe it.

David A. Lagnado, Department of Psychology, University College London, London; Steven Sloman, Department of Cognitive and Linguistic Sciences, Brown University.

Portions of the results of Experiment 1 were reported at the conference of the Cognitive Science Society, Fairfax, Virginia, August 2002. This work was funded by National Aeronautics and Space Administration Grant NCC2-1217.

We thank Michael Waldmann and Mark Buehner for many insightful comments on a draft of this article and David Shanks, Dave Sobel, and Mark Steyvers for helpful discussions.

Correspondence concerning this article should be addressed to David A. Lagnado, Department of Psychology, University College London, Gower Street, London WC1E 6BT. E-mail: d.lagnado@ucl.ac.uk

The current study aims to answer three questions. Is there an advantage of intervention over observation in causal structure learning? If so, what drives this advantage? Finally, what light does this shed on the cognitive mechanisms that people use to learn about causal structure?

Previous Research in Causal Learning

Although the distinction between observation and intervention is marked in theories of causal learning by the contrast between classical Pavlovian and instrumental conditioning (Mackintosh & Dickinson, 1979), differences with respect to the ease or manner of learning have not been thoroughly explored (for recent exceptions see Gopnik et al., 2004; Sobel, 2003; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). This is because previous research (see Shanks, in press, for an overview) has tended to presort the learner's environment into putative causes and effects and has focused on people's estimates of the strength of the causal relation between them[1] (cf. Tenenbaum & Griffiths, 2001).

Such learning situations minimize the disparity between observation and intervention because there is little difference between observing that a putative cause is present and intervening to make it present if you already know a causal relation exists. What matters in such a case is the degree to which this cause, once present, produces (or predicts) the effect. However, with more complex environments, or when variables are not prepackaged as causes or effects, the difference between observation and intervention may be critical to uncovering the correct causal structure.

Computational Models of Structure Learning

A formal framework for causal inference based on causal Bayes networks has been recently developed (Pearl, 2000; Spirtes, Glymour, & Schienes, 1993). This approach clarifies the relation between the probabilistic dependencies present in a set of data and the causal models that could have generated that data. It explicates the conditions under which it is possible to move from data to causal model (and vice versa) and permits the modeling of interventions on a causal system. We give an informal introduction to some of these ideas in a section below.

On the basis of this framework, computational methods have been developed for inferring causal structure from statistical data. There are two main approaches in the literature here: bottom-up constraint-based methods (e.g., Spirtes et al., 1993) and top-down Bayesian methods (e.g., Heckerman, Meek, & Cooper, 1999). Applied to human causal induction, the bottom-up approach supposes that people encode information about the probabilistic dependencies present in the experienced data and use this information to generate Markov equivalent causal models. In contrast, top-down approaches suppose that people use prior knowledge to select an initial model (or small set of candidate models) and then update these as they experience the statistical data.

A psychological version of the top-down approach was advanced by Waldmann (1996) and was termed the causal model theory. Waldmann argued that people's prior knowledge and assumptions not only provide initial causal models but also shape how the subsequent learning data are interpreted. In support of these claims, various experiments have shown that abstract knowledge about causality (such as causal directionality) can affect final causal judgments, even when the learning input is held constant (Waldmann, 1996, 2001; Waldmann & Hagmayer, 2001).

Most of these experiments have focused on people's causal strength estimates rather than looking directly at judgments about causal structure. A secondary aim of the current research was to produce empirical data that discriminates between the causal model and constraint-based approaches to structure learning.

Recent Work on Structure Learning

Steyvers et al. (2003) argued for a top-down Bayesian model of causal induction. In common with the research aims in this article, they too examined the contrast between observational and interventional learning, albeit within a different experimental paradigm. They used a novel task in which people infer the communication network between a group of alien mind readers. In an observation phase, participants see several trials, each depicting a specific pattern of communications between three aliens. They then select the network that best explains these data. This observation phase is followed by an intervention phase, in which participants make a single intervention (by implanting a distinctive word in a selected alien's head) and view several examples of the communication patterns that result. Finally, a second choice as to the correct network is made. Overall, participants improved their initial model choices once they had seen the results of this single intervention.

The paradigm adopted by Steyvers et al. (2003) differs from the standard causal learning paradigm in a variety of ways. In their set-up, category variables can take on a large number of values (different possible words) rather than being binary. This facilitates the detection of correlations and reduces the chances of mistaking a spurious correlation for a genuine one. They also use a relatively small number of trials with no corrective feedback. This reduces the ability of data-driven mechanisms to learn.

The experimental paradigm adopted in the present study is much closer to the standard causal learning paradigm. We used binary variables, a larger number of trials, and corrective feedback on each trial. We also introduced separate observation and intervention conditions with the same number of trials in either condition. Thus, people in the intervention phase made multiple interventions, which gave them the opportunity to experiment repeatedly. This design equalizes the number of trials in both the intervention and observation conditions, whereas in the design of Steyvers et al. (2003), the observation phase always precedes the intervention phase; thus, model choices based on observation alone are always made on less information than choices made after intervention. Another difference between these task paradigms is that our task marks out one variable as an outcome (in common with the standard paradigm), whereas in the mind-reading task, the outcome variable has no unique status.

We see these different task paradigms as complementary routes to the analysis of causal inference. They may also tap into different kinds of learning processes. This issue will be discussed in the General Discussion section.

[1] In such experiments, participants can judge that there is no causal link between a candidate cause and the effect by giving a causal strength estimate of zero.

Three Possible Reasons for the Advantage of Intervention

Learning through experiment might be more effective than observation for a variety of reasons. We focused on three, as follows.

Information

The two forms of learning can lead to distinct bodies of information. Observational learning involves the sampling of a system's autonomous behavior: One is exposed to patterns of data (e.g., drug-use rates and crime rates across various time periods or districts) generated by the system's underlying causal model. In contrast, with interventional learning patterns of data are observed conditional on specific actions on the system (e.g., the police observe what effect cracking down on drug usage has on the crime rate). One consequence of this difference is that interveners are able to modulate what type of data they observe. They can select the variety of information they see and hence concentrate on particular subsets of the system's behavior. For example, in exploring a software package, an individual can focus exclusively on a small subset of functions and master these before moving onto more complex procedures. Note that only in the observational case is it possible to see a representative sample of the system as it operates autonomously. We term this a *selection-based difference* in information.

Another consequence of intervention is that it can lead to the modification of the causal model that generates the data. This occurs whenever the system variable intervened on has other causal links feeding into it. If the intervention is successful, then the value of this variable will no longer be affected by these other causal mechanisms because their effects are being overridden by the intervention. In effect, a new causal model is created with those links removed.

To illustrate, consider the causal model depicted in the upper panel of Figure 1, which shows a simplified model of a battery-driven torch. There are three binary variables, battery, current flow, light, each of which can take the values *on* or *off*. Under normal operation turning the battery on causes the current to flow, and this in turn causes the light to be on.

Suppose that one intervenes on this system by stopping the current flow (e.g., by breaking the circuit). This can be represented by setting the current flow variable to off and by removing the causal link from battery to current flow because battery no longer has a causal effect on current flow—only the intervention does. The modified model is shown in the lower panel of Figure 1. It represents the intuition that if one intervenes to stop the current flow, one cannot draw any new conclusions about the status of the battery.[2] One can, however, expect the light to go off, because the causal link from current flow to light remains intact (for full details see Pearl, 2000; Spirtes, Glymour, & Schienes, 1993).

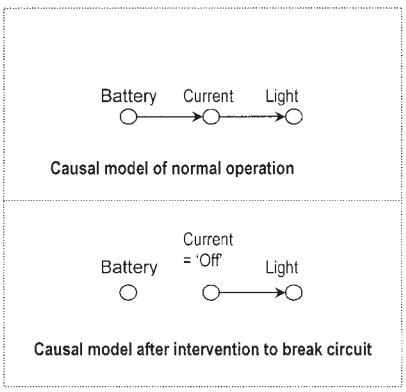

Figure 1. Simplified causal model of a battery-driven torch and after an intervention to break the circuit.

It is clear that the kind of information that results from such modifications to a causal system is peculiar to intervention and not observation. We term this a *modification-based difference* in information. For the purpose of learning causal structure, this can have several advantages. First, as Mill (1843/1950) pointed out, by performing an experiment, we can discriminate causal structures that are impossible to distinguish through observation alone. We have already seen this in the earlier example, in which an intervention on drug usage permitted the inference that a high incidence of drug usage leads to higher crime rates. The same principle applies with more complex causal structures. For instance, consider the task of distinguishing between a causal chain model ($A \rightarrow B \rightarrow C$) and a common cause model ($A \leftarrow B \rightarrow C$). For concreteness, suppose we have two competing models of the relationship between anxiety, insomnia, and tiredness. According to the chain model, anxiety causes insomnia, and insomnia causes tiredness (without any other causal links between these three variables). According to the common cause model, insomnia is an independent cause of both anxiety and tiredness. These two alternative models are shown in the upper panel of Figure 2.

If you can only observe the data generated by the system (e.g., by looking at a database of people visiting a sleep clinic, where for each patient there is a chart listing levels of anxiety, insomnia, and tiredness) and have no other clues about causal structure, it is impossible to determine whether the underlying causal structure is a chain or a common cause. This is because the two structures generate data that share the same conditional and unconditional dependencies. All three variables are unconditionally dependent, but anxiety and tiredness are independent conditional on insomnia. That is, the probability of tiredness given insomnia is the same as the probability of tiredness given insomnia and anxiety. This corresponds to the fact that in both models insomnia "screens-off" anxiety from tiredness. If you know that someone has insomnia, the additional fact that they have anxiety does not alter the probability that they are tired. In other words, there is no direct causal link between anxiety and tiredness; any covariation between them is mediated by insomnia.

Models that share the same conditional and unconditional dependencies are termed *Markov equivalent* (Pearl, 2000; Spirtes et al., 1993). An appropriate intervention, however, enables the two structures to be distinguished. If one gives some of these patients a drug that promotes good sleep (and is known to have no side-effects on anxiety level), we are effectively intervening to set the insomnia variable to *off*. This is a critical test. If we observe that the levels of anxiety for these patients tend to drop, we can conclude that the common cause model is correct; that insomnia

[2] The base rate probability of the battery being on is the same as in the original unmodified model.

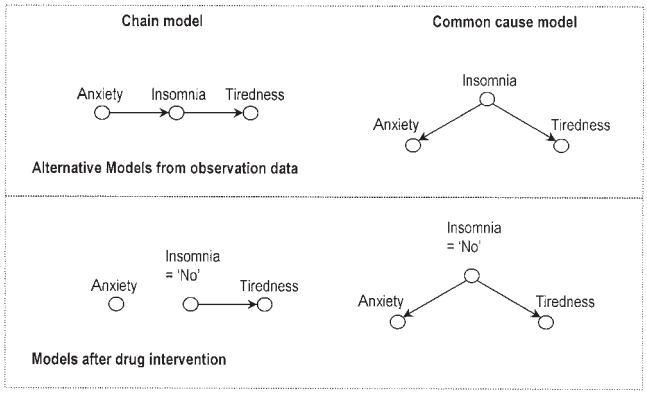

Figure 2. Distinguishing between two alternative causal models by means of an intervention.

causes anxiety. If we observe that they still exhibit high levels of anxiety, we can conclude that the chain is correct; that anxiety causes insomnia and not vice versa. The two new models created by the drug intervention are shown in the lower panel of Figure 2.

A related advantage of intervention is that it allows us to isolate and test specific subsystems and hence enables us to perform unconfounded experiments. This applies both to standard experimental practices (e.g., the use of randomizing procedures) and to the informal experiments we conduct in our day-to-day life. By performing a variety of sporadic actions, we reduce the chances that their consequences are in fact caused by some external factor that just happens to covary with these actions.

Two assumptions underlie modification-based learning. First, interventions are made relative to a specific causal model; thus, for someone to learn about causal structure through intervention, they must have a prior set of candidate causal models. These might be induced on the basis of observational data or through prior knowledge or assumptions about the causal system in question. Second, in common with observationally based learning, structural inferences can only be made on the assumption that there are no unobserved common causes (this is known as the *sufficiency assumption*). Without this assumption, a specific intervention may fail to discriminate between candidate causal structures. For example, if there is a hidden common cause of anxiety and tiredness (such as a mysterious virus), an intervention on insomnia will not break the dependence between these two variables.

Decision Demand

Quite aside from the informational differences between experiment and observation, the tasks also differ in terms of their demands. In particular, when experimenting, people must decide what intervention to make, whereas when observing, people are not obliged to make such a decision. As a consequence, it is possible that interveners become more engaged in the learning task and thus outperform observers. This possibility will be discussed in more detail in the introduction to Experiment 3.

Temporal Cue

Another important difference between intervention and observation is that in the former, one is customarily supplied with an implicit temporal cue. This is because our experience of our own actions precedes our experience of the effects of these actions, and this maps onto the fact that actions never occur after their effects. In other words, if you perform an action, you can be reasonably sure that any subsequent changes are effects rather than causes of this action (both your action and the subsequent changes may be common effects of a confounding cause, but this possibility can be ruled out by repeated actions performed independently of other variables in the system). Thus, your intervention itself furnishes you with a temporal cue relevant to uncovering causal structure: Changes that happen subsequent to your intervention are very likely to be effects of your action and cannot be causes of it.

However, in the case of observation, if one observes data simultaneously (e.g., when consulting a chart summarizing the patient's levels of anxiety, insomnia, and tiredness) no temporal cue is available. If the receipt of data does contain temporal delays, then these can provide some cue to causal structure, but they will be less stable indicators than in the case of intervention, because one sometimes receives information about effects before one receives information about their causes, for example, when symptoms are used to diagnose a disease or when the causes of an airplane crash are inferred from its black box recording. This temporal priority cue is discussed in more detail in the introduction to Experiment 4.

In the four studies reported in this article, we aimed to explore how these various factors—information (selection-based and modification-based), decision demand, and temporal cues—affect the difference between interventional and observational learning. In the first experiment, we established that there is indeed a difference between people's causal model selections across the two types of learning. In the second experiment, we examined whether this is due to a modification-based difference in information. In the third experiment, we looked at both selection-based differences in information and decision demand, whereas in the fourth experiment, we focused on the effect of temporal cues.

Experiment 1

The central aim of this experiment was to compare the observational and interventional learning of a simple causal model. We used a standard observational learning paradigm (e.g., Shanks, 1995, in press) but adapted it to include an interventional learning condition and a model selection question. The learning data were generated from a simple three-variable chain model, and performance was assessed both through a model selection question and a set of conditional probability judgments.

Our reasons for using a causal chain structure were two-fold. First, although such a structure is relatively simple and pervasive in many everyday causal environments, it has not been investigated much in standard causal learning experiments (for recent exceptions, see Ahn & Dennis, 2000; Waldmann & Hagmayer, 2001). Second, it is the simplest structure that can generate qualitatively different data according to whether one observes or intervenes. As noted in the introduction, an intervention on the intermediate variable of a three-variable chain will modify its causal structure, rendering the intervened-on variable independent of its usual causes. This contrasts with the observational case in which the causal system remains unmodified. By using a chain structure in our experiments, we can explore whether people are

sensitive to such structural modification and whether they can use it to guide their learning of causal structure.

Method

Participants and apparatus. Thirty-six undergraduates from Brown University received course credit for their participation. All participants were tested individually. The entire experiment was run on a personal computer.

Materials. The presentation of on-screen information was very similar in both the observational and interventional tasks. For both tasks, two different cover stories were used. In the chemist scenario, the two potential causes were acid level and ester level, and the outcome was the production of a perfume. In the space engineer scenario, the two potential causes were temperature level (of the oxidizing agent) and pressure level (in the combustion chamber), and the outcome was the launch of a rocket. For illustrative purposes, we discuss examples using just the chemist scenario, but both scenarios were used throughout all the experiments.

In the learning phase, labels for the two potential causes were presented side-by-side, each with its own visual icon above and each with two boxes labeled *HIGH* and *LOW* directly underneath. A screen-shot of the learning phase is shown in Figure 3. On each trial in the observational condition, these boxes were highlighted in blue to display the appropriate test result. On each trial in the interventional condition, participants could click on one of these four boxes to set one variable to either *HIGH* or *LOW*. In both conditions, the label and visual icon for the outcome variable were displayed to the right of the cause variables. There were also two boxes directly underneath (labeled *PRESENT* and *ABSENT* in the chemist scenario and *LAUNCH* and *FAIL* in the space engineer scenario), one of which was highlighted in blue depending on the trial outcome.

The learning set in both conditions was constructed on the basis of the chain model shown in Figure 4. In the observation condition, this resulted in each participant seeing the same distribution of trials (shown in Table 1) but with order randomized. In the intervention condition, the make-up of each trial depended on which intervention was made. The values for the variable that was not intervened on, and for the outcome variable, were generated at random according to a suitably modified chain model. For example, if a participant intervened to set the source variable (e.g., acid) to high, the intermediate variable (ester) was made high, with a probability equal to 0.8 and conditional on ester being high the outcome variable (perfume) was made present with probability equal to 0.8. However, if someone set the intermediate variable (e.g., ester in the chemist scenario) to high, the other cause variable (acid) was made high, with a probability equal to 0.5 (its base rate of occurrence), and the outcome variable was made present with a probability equal to 0.8.

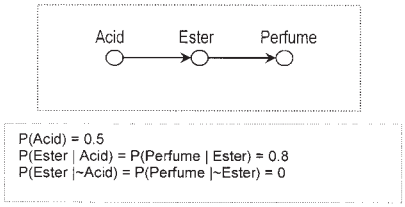

Figure 4. Causal graph used to generate stimuli for both observational and interventional tasks in Experiments 1 and 2.

The test phase was made up of two separate components: a model selection question and a set of probability judgments. For model selection, five candidate models were presented on the screen simultaneously (as shown in Figure 5), with an option box alongside each one. The probability judgments consisted of 12 questions. Eight of these asked for the conditional probability of the outcome variable (e.g., the probability of the perfume being produced) given a specified set of values for the two cause variables (e.g., given that the acid is high, and the ester is low). The other four questions asked for the probability of one of the cause variables, conditional on a specified value for the other one (e.g., the probability that the ester is high, given that the acid is low).

These questions were presented on the screen in a similar fashion to the learning phase: Visual icons and labels for the two cause variables were placed side-by-side, each with two boxes underneath indicating the values for each test question. A particular value was indicated with a blue highlight, and when the value was unknown, a question mark was placed

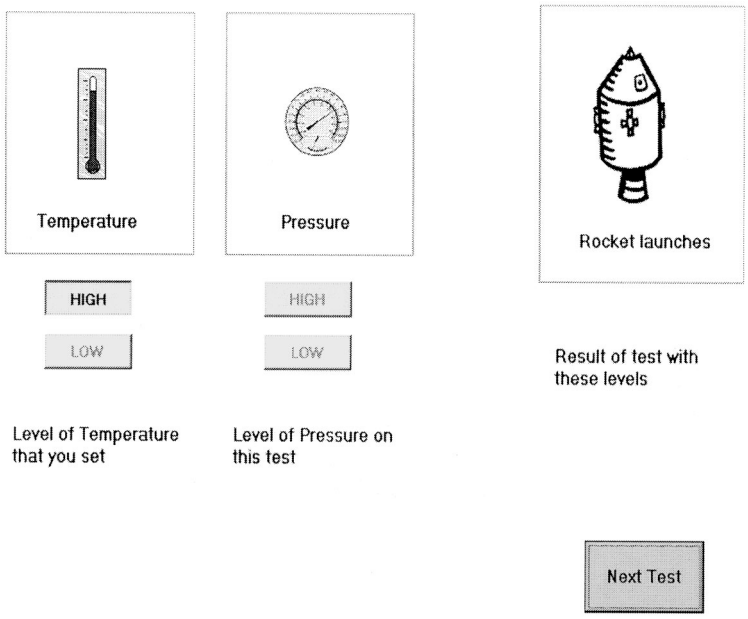

Figure 3. Screen-shot of learning phase for space engineer scenario.

over the appropriate pair of boxes. There was a 0–100 rating scale on which participants could register their probability judgment.

Design. The main factor (interventional vs. observational learning) was within-subject, with order and scenario (chemist or space engineer) counterbalanced. The order of the test phase components (model selection or probability judgments) was also counterbalanced, but the individual probability judgments were presented in a fixed order.

Procedure. Initial instructions to the participants included an introduction to the notion of a causal model with examples of the three types of models (chain, common cause, common effect) that they would be asked to choose between. Each participant then completed both an observational and an interventional learning task. Two cover stories were used, one for each task. Participants were asked to imagine that they were chemists (space engineers) running tests on a new perfume (rocket) in order to discover the underlying causal structure. They were told that previous tests had identified two variables as relevant to the success of the test. In the chemist scenario the variables were acid level (either high or low) and ester level (either high or low), and the outcome variable was whether the perfume was produced. In the space engineer scenario, the relevant variables were the temperature level of the oxidizing agent (either high or low) and the pressure level in the combustion chamber (either high or low), and the outcome variable was whether the rocket launched.

In the observation task, participants viewed the results of 50 test trials. On each trial they clicked on a button to view the values of the two cause variables (e.g., acid is high; ester is low) and whether the outcome occurred (e.g., the perfume is produced). All variable values were displayed simultaneously. The learning set was the same for each participant except for the order of presentation, which was randomized across participants. After viewing all of the trials, participants proceeded to the test phase.

In the learning phase of the intervention task, participants were able to set the value of one of the two cause variables. They then viewed the resulting values for the outcome variable and for the cause variable they had not intervened on. In contrast to the observation condition, the value for the intervened-on variable was viewed before the values for the two other variables. After running 50 tests, participants proceeded to a test phase identical to that of the observation condition.

There were two components to the test phase: a model selection question and a set of 12 probability judgments. In the model selection question, the five candidate models were presented on the screen simultaneously, and participants were asked to select the model that they believed was most likely to have produced the data that they had just seen. They registered this choice by clicking on the appropriate option button. Once a model was selected, a rating scale from 0–100 appeared with which participants expressed their confidence that this selection was correct. They could then select a second model if they wished or exit from the test. This process was repeated until they were satisfied that no other model was likely to have produced the data. In the probability judgment phase, the 12 questions were presented sequentially, and participants used the 0–100 rating scale to register their judgments.

Table 1
Frequency of Presented Instances in Observational Learning Condition in Experiment 1

Acid level	Ester level	Perfume	Frequency
High	High	Yes	16
High	High	No	4
High	Low	Yes	0
High	Low	No	5
Low	High	Yes	0
Low	High	No	0
Low	Low	Yes	0
Low	Low	No	25

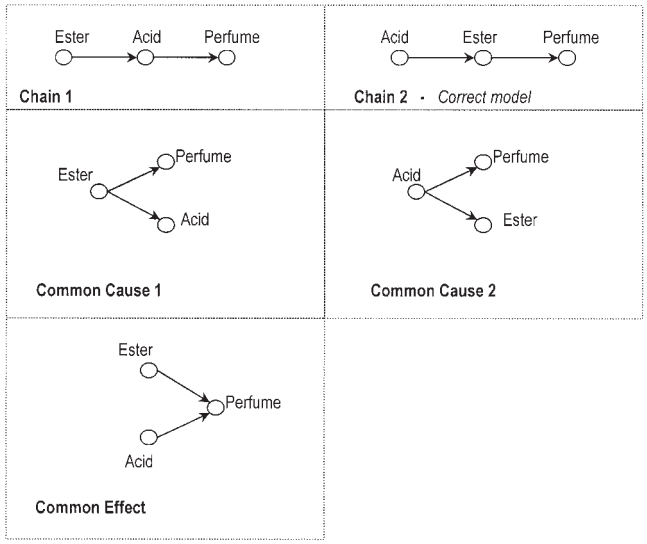

Figure 5. The five candidate causal models in the selection task.

After completing the first task, participants in the observation condition switched to an interventional learning task and vice versa. Each participant thus completed both an observational and interventional task.

Results

Model selection. The results for the model selection task are shown in Figure 6, with the correct chain model designated as Chain 2. We summed over the two scenarios because scenario had no significant effect on participants' choices. As can be seen from Figure 6, in both conditions only a minority of participants managed to select the correct chain model. However, the proportion of participants who chose the correct chain in the intervention condition (12 of 36) was greater than chance, $\chi^2(1, N = 36) = 4.1$, $p < .05$, whereas the proportion in the observation condition (5 of 36) did not differ from chance, $\chi^2(1, N = 36) = 0.84$, *ns*.[3] Furthermore, correct chain model selections were significantly higher in the intervention than in the observation condition, $t(35) = 2.02$, $p < .05$, one-tailed. There was also a strong bias in favor of the common effect model in the observation condition (67%, 24 out of 36), significantly greater than in the intervention condition (22%, 8 of 36), $t(35) = 4.39$, $p < .01$.

Judgments of conditional independence. On the chain model used to generate the learning data, acid (*A*) is independent of perfume (*P*) conditional on ester (*E*); in other words, ester screens-off acid from perfume. Sensitivity to this relation can be assessed by seeing whether participants' probability judgments conform to

[3] The chain model used to generate the data is Markov equivalent to the Common Cause Model 2. However, although not inconsistent with the observational data, this model requires an idiosyncratic parameterization whereby one effect (acid) occurs more often than its sole cause (ester). Only one person chose this model in the observation condition. Moreover, if this model is counted as a correct choice, the proportion of participants in the observation condition who select a correct model (6 of 36) is significantly less than chance, $\chi^2(1, N = 36) = 8.17$, $p < .01$. This is due to the strong bias for the common effect model.

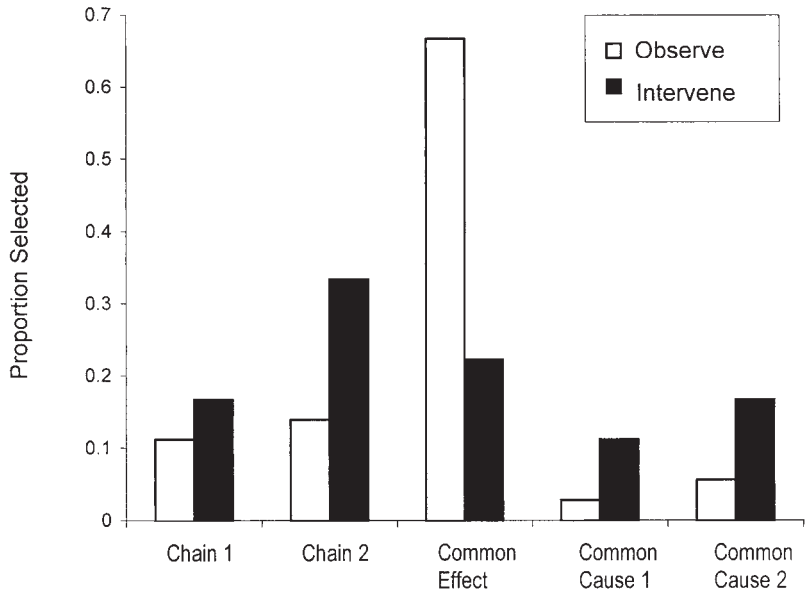

Figure 6. Model selection in Experiment 1 for interventional and observational conditions (the correct model is Chain 2).

the equality $P(P|A\&E) = P(P|E)$. The mean ratings for these two probabilities are shown in Figure 7. No significant difference was obtained between the two probabilities in the intervention condition; mean judged probability for $P(P|A\&E)$ was 81.6, for $P(P|E)$ it was 79.7, $t(35) = 0.78$, *ns*, $(1 - \beta = 0.987)$, suggesting that participants were sensitive to the conditional independence. This is reinforced by the fact that 21 of 36 (58%) participants judged the two probabilities equal. This contrasts with the observation condition, in which the mean probabilities differed substantially; mean judged probability for $P(P|A\&E)$ was 86.5, for $P(P|E)$ it was 63.2, $t(35) = 5.89$, $p < .0001$, and only 8 out of 36 (22%) participants judged them equal. Comparing the two conditions, a significantly greater number of participants obeyed screening-off in the interventional condition, $t(35) = 4.45$, $p < .001$. In short, participants tended to obey screening-off in the intervention but not in the observation condition.

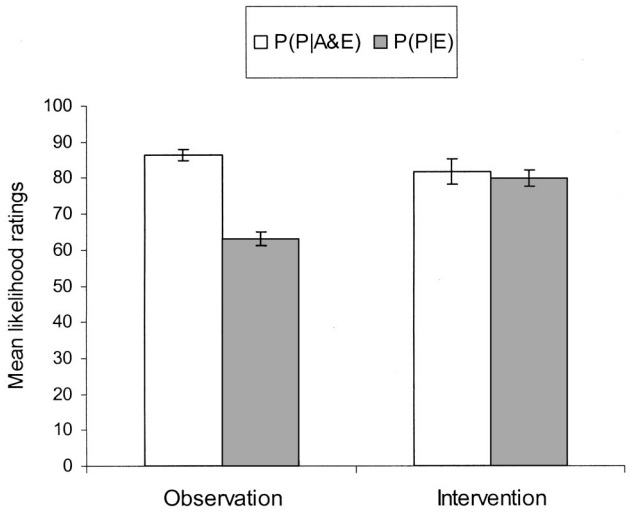

Figure 7. Assessment of screening-off: Mean conditional likelihood ratings for the outcome variable in Experiment 1. P = perfume; A = acid; E = ester.

Comparing model selections with probability judgments. Did people who chose the correct chain model also make probability judgments that conformed to the appropriate screening-off relation, and did those who chose the common effect model make judgments in conformity with the relations encoded by that model? There were no significant differences across observational and interventional conditions; thus, we report the percentages collapsed across conditions. Of those participants who did choose the correct chain model, 80% made probability judgments that obeyed screening-off (e.g., that acid is independent of perfume conditional on ester). Of those participants who chose the common effect, 82% made judgments that violated screening-off, and 64% made judgments that implied that the two potential causes (e.g., acid and ester) were independent. Both of these sets of judgments are consistent with the common effect model and not the chain model. Thus, in both observational and interventional conditions, participants' probability judgments tended to map onto their model selections, whether these were for the correct chain or the incorrect common effect.

Derived judgments of contingency. A common index for the degree to which two events are related is given by the statistical contingency measure $\triangle P$ (Allan, 1980). For two events, A and B, the contingency between A and B is determined by the difference between the probability of B given A and the probability of B given not-A. That is, $\triangle P_{AB} = P(B|A) - P(B|{\sim}A)$. Note that contingency is directional; thus, $\triangle P_{AB}$ is not necessarily equal to $\triangle P_{BA}$. Although not always an appropriate index of the causal relation between A and B (e.g., see Cheng, 1997; Pearl, 2000; Shanks, in press), a high degree of contingency is often suggestive of a causal

relation. In the current experiment, the programmed contingency between acid and ester ($\triangle P_{AE}$) is 0.8, as is the contingency between ester and perfume ($\triangle P_{EP}$). These contingencies provide a measure of the strength of these individual links. By deriving estimates of these contingencies from participant's probability judgments, we obtained an indirect measure of their sensitivity to the strength of these relations.

The derived $\triangle P$ between two variables A and B is given by:

$$\text{derived } \triangle P_{AB} = \text{judged } P(B|A) - \text{judged } P(B|{\sim}A).$$

This was computed for both the contingency between acid and ester ($\triangle P_{AE}$) and that between ester and perfume ($\triangle P_{EP}$), corresponding to the two directional links in the chain model. Again, observational and interventional groups did not differ; thus we collapsed across them. The most notable finding was that the judged contingency between ester and perfume, derived $\triangle P_{EP} = 0.64$, was much higher than that between acid and ester, derived $\triangle P_{AE} = 0.25$, $t(70) = 9.79$, $p < .0001$, although both had equivalent strength in the learning data (actual $\triangle P = 0.8$).[4] This underweighting of the contingency between acid and ester fits with the general failure to identify the causal link from acid to ester.

Discussion

In sum, the learning of causal structure improved with intervention, both with respect to number of correct model selections and sensitivity to the appropriate conditional independencies. However, the majority of participants in both conditions still failed to uncover the correct model and consistently underweighted the link between putative causes. This is most apparent in the observation condition, in which most participants selected a common effect model in which putative causes are independent. This tendency resonates with research in multiple-cue probability judgment (Hammond, 1996; Hastie & Dawes, 2001), in which models that assume the independence of causes fit the human data well. Indeed, most of the research in that paradigm operates on the assumption that putative causes are independent.

Experiment 2

What is it that drives the advantage for intervention over observation? Although the same underlying causal chain was used to generate the learning data in both the observational and interventional conditions in Experiment 1, participants did not receive the same information in both conditions. One difference is modification-based: intervening on the intermediate variable in a chain (e.g., ester in the chemist scenario) disconnects it from the source variable (acid). As a result of this intervention, a participant may create a trial in which the ester is high, and the perfume is produced, but the acid is low (indeed such a trial will occur on 40% of the occasions when an intervener sets ester to high). In contrast, an observer will never see a trial with this configuration of values; whenever acid is low, ester will also be low, and the perfume will not be produced. A second difference is selection-based: Interventions determine the frequencies of the different kinds of trials experienced. Therefore, interveners can choose a different distribution of trials than the representative set seen by observers. For instance, interveners can choose to see a predominance of trials in which a particular variable is high by repeatedly setting that variable high.

To equalize information across observational and interventional conditions, we modified the observation condition; thus, participants observed the results of another participant's interventions. Participants in this observation-of-intervention condition were told the precise nature of the intervention made on each trial (e.g., the variable intervened-on and the value set). In this way, participants in both conditions experienced identical learning sets.

If the advantage of intervention is based purely on the distinctive data that it generates, then there should be no difference between these two conditions. Unlike the contrast between observation and intervention in Experiment 1, in the current set-up the statistical data are equivalent in both conditions. From a statistical viewpoint, the observation-of-intervention condition is equivalent to the intervention condition. However, if an advantage for intervention persists, then an alternative factor will be implicated.

Method

Participants and apparatus. Twenty-two undergraduates from Brown University received $7 each for their participation. None had taken part in Experiment 1. All participants were tested individually, and the entire experiment was run on a personal computer.

Procedure and materials. The introductory instructions for all participants were identical to those in Experiment 1. All participants completed both an intervention and an observation-of-intervention condition (order counterbalanced). The intervention condition was identical to that in Experiment 1. In the observation-of-intervention condition participants were asked to imagine that they were laboratory assistants observing experimental tests conducted by a research chemist (or space engineer). On each of 50 trials, they clicked on just one button to view the value the chemist (engineer) had set for a particular variable (e.g., setting the acid level to high), the value the other variable took (e.g., ester level is low), and whether the outcome occurred (e.g., whether the perfume is produced). Thus, on each trial, all three variable values were displayed simultaneously, just as in the observation condition in Experiment 1.

To ensure that participants in this condition were aware that they were observing the results of someone else's interventions, the intervened-on variables on each trial were clearly highlighted and labeled as interventions (e.g., the label "The chemist set the acid to this level" appeared above the acid icon). Similarly, variables that were not intervened-on were also clearly labeled (e.g., the label "The level of ester observed on this trial" appeared above the ester icon).

The learning set for each participant was yoked to the data generated by a previous participant in the intervention condition. Thus, information was equalized across both interventional and observational conditions. Clearly, the information that each participant observed varied according to the pattern of interventions made by the participant who they are yoked to, although there was not too much variation in overall frequencies of tests conducted between different interveners. Table 2 shows the mean frequencies of trials that were generated by the interveners and thus viewed in the observation-of-intervention condition. Otherwise, the method was identical to that of Experiment 1.

[4] In the observation condition, the experienced contingencies for all participants were exactly equal to the programmed contingencies. In the intervention condition, these were only approximately equal because participants controlled the trial frequencies they experienced. However, the discrepancies between experienced and programmed contingencies were very small (e.g., mean experienced $\triangle P_{EP} = 0.79$; $\triangle P_{AE} = 0.81$).

Table 2
Mean Frequencies (Across All Participants) of Trial Types Generated by Interveners in Experiment 2

Intervention	Mean frequency	Acid level = high	Ester level = high	Perfume = present
Set acid high	14.5	14.5	11.4	9.2
Set acid low	11.1	0	0	0
Set ester high	15.5	7.5	15.5	12.5
Set ester low	8.9	4.5	0	0

Results

Model selection. The results for the model selection task are shown in Figure 8, with the correct chain model again designated as Chain 2.[5] Replicating the results of Experiment 1, the proportion of participants who chose the correct chain in the intervention condition (12 of 22) was significantly greater than chance, $\chi^2(1, N = 22) = 6.01$, $p < .05$, whereas the proportion in the observation-of-intervention condition (4 of 22) was not, $\chi^2(1, N = 22) = 0.56$, *ns*. In correspondence, correct chain model selections were significantly higher in the intervention than in the observation-of-intervention condition, $t(21) = 3.46$, $p < .01$. There was also a strong bias in favor of the common effect model under observation of intervention (55%; 12 of 22) but not under intervention (18%; 4 of 22), which was a significant difference, $t(21) = 3.46$, $p < .01$.

Judgments of conditional independence. Participants' mean ratings for the two relevant probabilities are shown in Figure 9. As in Experiment 1, no significant difference was obtained between the two probabilities in the intervention condition, suggesting that participants were sensitive to screening-off. The mean judged probability for $P(P|A\&E)$ was 85.3 and for $P(P|E)$ it was 84.7, $t(21) = 0.20$, *ns*, $(1 - \beta = 0.900)$. This is reinforced by the finding that 13 of 22 (59%) participants judged the two probabilities equal. However, although the two mean probabilities differed in the observation-of-intervention condition ($P[P|A\&E] = 81.7$, $P[P|E] = 74.0$), this was only marginally significant, $t(21) = 1.81$, $p = .08$, and 8 of 22 (36%) participants judged them equal. Further, in a comparison between the two conditions, the number of interveners that obeyed screening-off was not significantly greater than the number of observers (sign test, $n = 11, p = .11$). This suggests that the informational content of the data generated by intervention (as opposed to pure observation) does promote participant's sensitivity to the screening-off relation. Thus, the wider variety of trials and the distinct pattern of data that results from interventions on the intermediate variable in the chain helps people to establish that the intermediate variable screens off the other two variables. However, this does not invariably lead to the selection of the correct model.

Comparing model selections with probability judgments. Observational and interventional conditions did not differ significantly, so we report the percentages collapsed across conditions. In line with the findings in Experiment 1, of those participants who chose the correct chain model, 75% made probability judgments that obeyed screening-off (e.g., that acid is independent of perfume conditional on ester). However, in contrast with Experiment 1, of those participants who chose the common effect model, only 50% made judgments that violated screening-off (in Experiment 1 it had been 82%), and 57% made judgments that implied that the two potential causes (e.g., acid and ester) were independent. Thus, in this experiment, those who chose the correct model made probability judgments consistent with that selection, but half of those who chose the common effect model made probability judgments that were consistent with the chain model and not the common effect model.

Derived judgments of contingency. As in Experiment 1, no significant difference was obtained between observational and interventional groups; thus, we collapsed across them. Once again, the most notable finding was that the judged contingency between the ester and perfume variables, derived $\triangle P_{EP} = 0.71$, was much higher than that between acid and ester, derived $\triangle P_{AE} = 0.32$, $t(44) = 6.90$, $p < .0001$, even though both had equivalent strength in the learning data (actual $\triangle P = 0.8$).

Discussion

With respect to the model selection task, the results from this experiment closely parallel those of Experiment 1. There is a marked improvement in the intervention condition compared with the observation-of-intervention condition, which cannot be explained in terms of interveners receiving better information because the distributional content of the learning data was equated across conditions. However, with respect to sensitivity to the screening-off relation, interveners were not significantly better than observers.

Compared with Experiment 1, the intervention condition produced the same proportion of people obeying screening-off (58% as compared with 59%), but in the new observation (of intervention) condition there were 36% as opposed to 24% in the pure observation condition. In short, observers of intervention in this experiment were more sensitive to screening-off than the pure observers in Experiment 1, even though they were no better at selecting the correct model. Indeed, in this experiment 50% of the observers who obeyed screening-off selected the common effect model, for which the screening-off relation should not hold. This suggests a partial dissociation between whether people select the correct model and whether their judgments are sensitive to screening-off. The modification-based information afforded by intervention (or observation of intervention) serves to improve sensitivity to screening-off without guaranteeing the selection of the correct model. Finally, in accordance with the results in Experiment 1, in both conditions people's derived contingency judgments suggested an underweighting of the link between putative causes.

Experiment 3

Although the results of Experiment 2 suggest that neither modification-based nor selection-based information alone is driving the advantage of intervention, the selection-based account could be adjusted to accommodate our findings. By having observers watch another person's interventions, we denied them the chance to conduct their own brand of hypothesis testing. In effect they were viewing someone else's selections, and this may be very different from being able to make their own.

[5] Because both conditions involved interventions on the chain model, there was only one uniquely correct model.

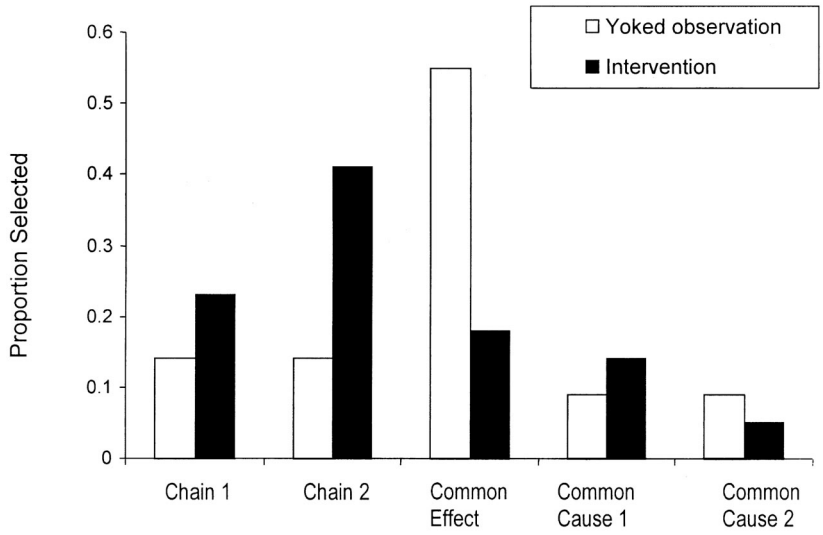

Figure 8. Model selection in Experiment 2 for interventional and yoked observational conditions (the correct model is Chain 2).

Relatedly, because the observers in Experiments 1 and 2 did not have to make any selections themselves, they did not need to engage in any decision making. In contrast, interveners had to choose what intervention to make on each trial. In the introduction, we termed this a *decision demand* and suggested that it might increase focus on the task and thus enhance learning.

To investigate these possibilities we introduced two new conditions—*selected observation* and *forced intervention*. In selected observation, participants had to actively select a particular value for one of the variables (e.g., to choose to look at a trial on which acid is set high) prior to viewing the values that the other variables take on that trial. This parallels the task confronting people in the intervention condition, insofar as it requires the active selection of one of the variable values and is thus likely to recruit similar decision or hypothesis testing processes. It differs, however, in that people are still just observing variable values rather than setting them. Thus, selecting a value for the intermediate variable in a chain (e.g., choosing to view a trial on which ester was low) does not involve modification of the causal model (e.g., severing the link from acid to ester).

In the forced intervention condition, the selection requirement is removed from the intervention task. In this condition, people are simply told which intervention to make; thus, the need to decide what intervention to make is removed along with the opportunity to pursue a specific hypothesis-testing strategy.

We compared these two conditions with the two original observation and intervention conditions from Experiment 1. The former was termed *forced observation*, because it did not involve selection requirements; the latter was termed *selected intervention*, because it did not involve selecting an intervention on each trial. All four conditions are shown in Figure 10.

If the advantage of intervention is due to the freedom (or requirement) to make selections, then one would expect selected observation to improve performance relative to forced observation, and forced intervention to reduce performance relative to selected intervention.

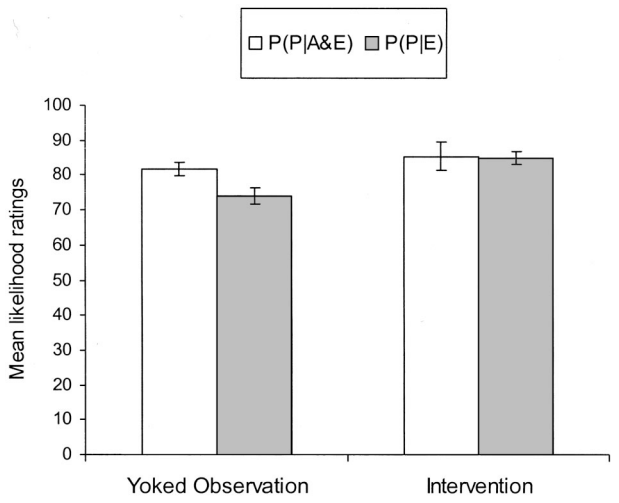

Figure 9. Assessment of screening-off: Mean conditional likelihood ratings for the outcome variable in Experiment 2. P = perfume; A = acid; E = ester.

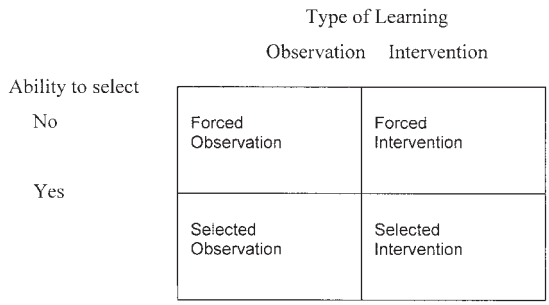

Figure 10. The four conditions in Experiment 3 resulting from crossing type of learning (observation vs. intervention) with ability to make selections (forced vs. selected).

In addition, in this experiment we modified the parameters of the chain model; thus, the learning data it generated were fully probabilistic. In the first two experiments, the model was semideterministic such that a cause could occur without its effect (e.g., $P[P|E] < 1$) but an effect could not occur without its cause (e.g., $P[P|{\sim}E] = 0$). One feature of such a semideterministic chain model is that it cannot be learned by constraint-based algorithms (e.g., TETRAD; see Spirtes et al., 1993) that infer graph structure from the determination of pairwise unconditional and conditional dependencies. This is because such an algorithm needs to compute all the relevant conditional independencies in the generated data, but some of these will be undefined when the learning set is semideterministic. More specifically, to induce the causal chain used in Experiments 1 and 2, an algorithm such as TETRAD needs to establish that low acid is independent of perfume conditional on high ester: $P(P|{\sim}A\&E) = P(P|E)$. However, no instances of low acid and high ester appear in the learning set, so the conditional $P(P|{\sim}A\&E)$ is undefined. If people also use constraint-based methods, as has been suggested by Glymour (2001) and Gopnik et al. (2004), then we have an explanation for their poor performance in the observation condition. They fail to induce the correct causal structure because they are unable to compute all the relevant conditional independencies. A simple test of this claim is to make the learning set fully probabilistic; thus, all the relevant conditional independencies are computable. On such a learning set TETRAD readily infers the correct chain structure. The empirical question is whether humans can too.

Finally, in order to give participants more flexibility in the causal models they could select, we used a new response method. Rather than choosing between five complete models, we allowed participants to build up the complete model by selecting individual causal links.

Method

Participants and apparatus. Forty-eight undergraduates from Brown University received $7 for their participation. None had taken part in previous experiments. All participants were tested individually, and the entire experiment was run on a personal computer.

Procedure and materials. The experiment had a mixed design, with type of learning (observation vs. intervention) as a between-subjects factor, and ability to make selections (forced vs. selected) as a within-subject factor. All participants received the same introductory instructions and cover stories (chemist and space engineer scenarios) as in previous experiments. Half of the participants then proceeded to a two-task observation condition, which consisted of both a forced observation task and a selected observation task (task order and cover story counterbalanced). The forced observation task was the same as the observational task in Experiment 1, in which participants were shown the values for all three variables simultaneously. The learning phase consisted of 50 trials, constructed according to a fully probabilistic version of the chain model from Experiments 1 and 2 (see Figure 11). The learning set was the same for each participant (shown in Table 3) but with order randomized.

In the selected observation condition, participants were asked to imagine that they were researchers consulting previous test reports that had been filed by the company secretary. On each trial they chose which kind of test report to look at by clicking on one of four options: "Look at a test with high acid level," "Look at a test with low acid level," "Look at a test with high ester level," and "Look at a test with low ester level." Having made their choice, they then clicked on another button marked "Check the other results in this test report" to see the values for the other two variables. These latter values were determined from a precompiled list of observation trials generated by the probabilistic chain model in Figure 11. Thus, participants in the selected observation condition experienced the same conditional probabilities as those in the forced observation condition. Appropriate labels were used throughout to remind participants that they were looking at the results of past test reports (minimizing the possibility that they misinterpret their selections as interventions). For example, below the icon for the ester variable was the label "The level of ester in this test report."

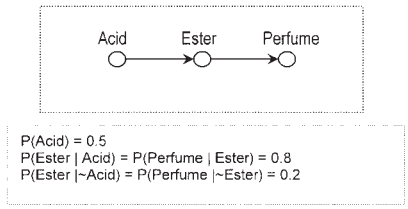

Figure 11. Causal graph used to generate stimuli in Experiment 3.

The other half of the participants completed a two-task intervention condition. One of these tasks—selected intervention—was the same as the intervention condition in Experiments 1 and 2. On each trial, participants selected a particular intervention and then viewed the values taken by the other two variables. The same fully probabilistic chain model was used, but values were computed according to the interventional calculus (e.g., intervening to set ester high did not affect the probability that acid was high, which remained at 50%). In the forced intervention task, participants were told that they were laboratory assistants carrying out the instructions of a superior. On each trial they received an instruction to set one of the potential cause variables to a particular value (e.g., on this trial set ester to low) and then viewed the values for the other two variables. The pattern of data they experienced (both which intervention they were instructed to make and the values taken by the other variables) was determined by the pattern of data generated by the previous participant in the selected intervention condition. Thus, the forced interveners were yoked with the interveners who could select their own interventions.

Each learning phase, regardless of condition, was followed by a similar test phase. This consisted of a set of conditional probability judgments (as in Experiments 1 and 2) plus a new causal model selection task. In the latter, participants were presented with a diagram with the three variables connected by faint arrows (see Figure 12) and had to select or deselect each individual causal link. Thus, for each link from *X* to *Y*, they were asked to click *yes* or *no* according to whether they believed that *X* causes *Y*. If they clicked *yes*, the arrow from *X* to *Y* was highlighted; if they clicked *no*, it disappeared.

Results

Model selection. The new response mode allowed a choice of 16 possible models; thus, random responding would lead to a choice expectation for each model of 0.0625. There was a wide range of models constructed in each of the four conditions, and only a minority of participants generated the correct chain model. Table 4 shows the proportions for the five most common models (these make up over 65% of the total choices). We conducted separate chi-square tests on each condition to see whether the frequency of correct chain models differed from chance. In the two conditions in which participants were free to make selections—selected intervention and selected observation—the proportion of correct chain models were the same (17%), and both were significantly greater than chance, $\chi^2(1, N = 24) = 4.44$, $p < .05$. In contrast, in the forced observation condition (0%) and the forced intervention condition (13%), neither proportion differed from chance (for both conditions: $\chi^2[1, N = 24] = 1.60$, *ns*).

This difference between selected and forced conditions can also

Table 3
Frequency of Presented Instances in Observational Learning Condition in Experiment 3

Acid level	Ester level	Perfume	Frequency
High	High	Yes	16
High	High	No	4
High	Low	Yes	1
High	Low	No	4
Low	High	Yes	4
Low	High	No	1
Low	Low	Yes	4
Low	Low	No	16

be seen in the percentages choosing the correct chain model: significantly more in selected observation (17%) than in forced observation (0%), $t(23) = 2.15$, $p < .05$. However, there was no corresponding difference between selected intervention (17%) and forced intervention (13%), $t(23) = 1.00$, *ns*.

Comparing the between-subjects conditions, there was a significant difference between forced intervention and forced observation, $t(46) = 1.81$, $p < .05$, one-tailed, but no difference between selected observation and selected intervention, $t(46) = 0.00$, *ns*.

The difference between selected and forced observation is the same if we include the Markov equivalent common cause model in our analysis. Recall that although the learning data were generated by a chain, this structure is Markov equivalent to the common cause; thus, it could also be counted as a correct choice. If we do count it as a correct choice, the proportion of participants who chose a correct model (chain or Markov equivalent common cause) in the selected observation condition (34%) was again greater than chance, $\chi^2(1, N = 24) = 9.52$, $p < .01$, whereas the proportion in the forced observation condition (4%) was not, $\chi^2(1, N = 24) = 1.52$, *ns*. Likewise, when comparing between conditions, there were significantly more correct choices in selected observation than in forced observation, $t(23) = 3.08$, $p < .01$.

Finally, the main choices made in the forced observation condition were either a single link model (38%) or the common effect (17%). Both of these were chosen significantly more often than in the selected observation condition: for the single link model, $t(23) = 2.16$, $p < .05$; for the common effect, $t(23) = 1.81$, $p < .05$, one-tailed.

Judgments of conditional independence. Recall that the correct chain model implies that ester screens-off acid from perfume, and thus $P(P|A\&E) = P(P|E)$. The mean ratings for these two probabilities in each of the four conditions are shown in Figure 13.

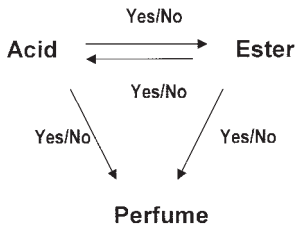

Figure 12. Causal model selection task for chemist scenario in Experiment 3.

Table 4
Proportion of Models Chosen in Each Condition in Experiment 3

Condition	a → e ↙ p: Correct chain	a ← e ↙ p: Common Cause 2	a ⇄ e ↙ p: Hybrid	a e ↘ ↙ p: Common effect	a e ↙ p: Single link
Forced observation	0.00	0.04	0.13	0.17	0.38
Selected observation	0.17	0.17	0.04	0.04	0.17
Selected intervention	0.17	0.17	0.04	0.17	0.13
Forced intervention	0.13	0.08	0.17	0.21	0.08

Note. a = acid; e = ester; p = perfume.

We conducted a mixed analysis of variance (ANOVA) with probability judgment ($P[P|A\&E]$ vs. $P[P|E]$) and ability to select (*yes* vs. *no*) as within-subject factors and type of learning (observation vs. intervention) as a between-subjects factor. There was a main effect of probability judgment, $F(1, 46) = 45.48$, $p < .001$, but no effect of ability to select, $F(1, 46) = 0.03$, *ns*, nor type of learning, $F(1, 46) = 0.65$, *ns*, and no interactions. Thus, in line with the low performance on the model construction task, the mean probability judgments for participants in all four conditions violated screening-off.

A different pattern emerges, however, if we separate out the probability judgments of those participants who gave a correct model (either the chain or the Markov equivalent common cause). These data are displayed in Figure 14 for observers and interveners. We grouped forced and selected conditions together because there were too few correct responses in forced observation to make meaningful comparisons across these two conditions. Using these probability judgments, we conducted a mixed ANOVA with probability judgment ($P[P|A\&E]$ vs. $P[P|E]$) as a within-subject factor and type of learning (observation vs. intervention) as a between-subjects factor. There was no main effect of probability judgment, $F(1, 14) = 0.015$, *ns* ($1 - \beta = 0.753$), nor of type of learning, $F(1, 14) = 0.931$, *ns*, and no interaction. This shows that the probability judgments for those participants who generated a correct model did obey screening-off, and that this held irrespective of whether they were intervening or observing. This is further confirmed by looking at individual responses: of those participants who chose the correct model, eight of nine observers (89%), and five of seven interveners (71%) obeyed the screening-off relation.

Derived judgments of contingency. As in Experiments 1 and 2, no significant difference was obtained between groups; thus, we collapsed across them. Once again, the most notable finding was that the judged contingency between the ester and perfume variables, derived $\triangle P_{EP} = 0.49$, was significantly higher than that between acid and ester, derived $\triangle P_{AE} = 0.27$, $t(94) = 6.33$, $p < .0001$, even though both had equivalent strength in the learning data (actual $\triangle P = 0.6$).

Discussion

The overall performance with a fully probabilistic environment was markedly lower than with the semiprobabilistic environment in Experiments 1 and 2. The additional noise made the task more difficult and did not, contra the prediction of a constraint-based

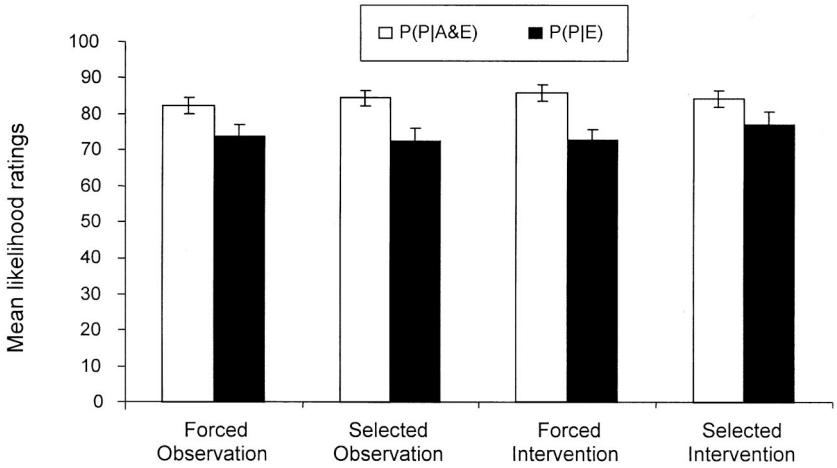

Figure 13. Assessment of screening-off: Mean conditional probability judgments for each condition in Experiment 3. P = perfume; A = acid; E = ester.

method such as TETRAD, enable people to compute the full set of conditional independencies or improve their model selections. The main reason for the increase in difficulty was the introduction of spontaneity — on each trial there was a 20% chance of a variable being high even though its immediate cause was low (e.g., when ester was low, perfume was produced with a 20% probability). This made it harder to establish that acid and perfume were independent given ester, because on some trials perfume would be produced when acid was high but ester was low. Such an arrangement was impossible in the semiprobabilistic environment in Experiments 1 and 2. Further, the potential benefits of interventional information were reduced because on some occasions an intervention to set ester low would result in perfume being produced and acid (coincidently) being high. Again, this was impossible in the previous experiments.

Despite the increase in task difficulty and the correspondent drop in performance, certain conclusions can be drawn from the data. With respect to correct model selections, selected observation, selected intervention, and forced intervention conditions did not differ, but in all three, learning was easier than in the forced observation condition. At first sight, the marked improvement for observers who were able to make their own selections compared with those who could not suggests an advantage due to either selection-based information or a decision demand. However, the fact that participants in the forced intervention condition performed no worse than in the selected intervention condition undermines both of these possibilities. If a selection-based difference or a decision demand was driving the advantage, then one would expect interveners who were not required to make any selections or decisions to perform worse than those who were.

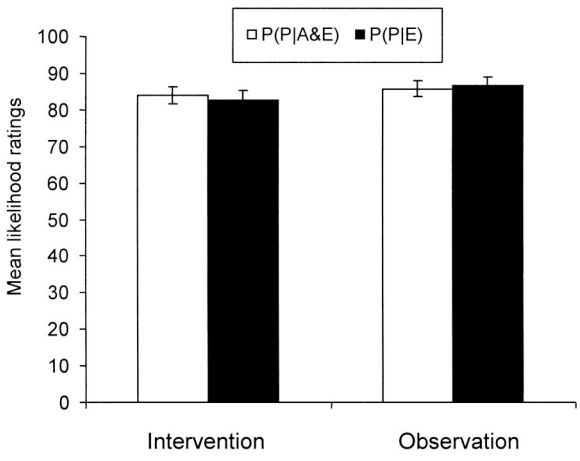

Figure 14. Assessment of screening-off: Mean conditional probability judgments for interveners and observers who chose correct model in Experiment 3. P = perfume; A = acid; E = ester.

One factor that does differentiate between selected observation, selected intervention, and forced intervention on the one hand and forced observation on the other, is the existence of a temporal delay for receipt of information in the former conditions. Thus, when participants have to make a selection or intervention themselves, or watch an intervention being made, they will know the value of the selected or intervened-on variable before the values of the other variables. As mentioned in the introduction, this can provide an important cue to the underlying causal structure, because interveners can assume that any subsequent changes are effects, not causes, of the intervened-on variable.

This would explain the advantage of both intervention conditions over the forced observation condition and also the finding that, relative to participants in the forced observation condition, those in selected observation were more likely to choose the correct chain or its Markov equivalent common cause (Model 5). Selecting the intermediate variable (e.g., ester) does not disconnect acid from ester; thus, the two variables are still highly correlated. If participants are using temporal delay as a cue to causal structure, they are likely to take this covariation as evidence in support of a causal link from ester to acid. This seems to be reflected in the results, in particular the finding that participants in selected observation endorsed the chain and common cause equally often and endorsed both more often than those in forced observation.

Furthermore an explanation in terms of the temporal cue hypothesis would help explain why people find the task so difficult, even in the intervention condition. In our experimental set-up, an intervention on the intermediate variable in a chain (e.g., ester) disconnects the link from acid to ester, and the value of acid is determined by its base rate probability. Because this base rate is 50%, establishing that, on trials on which ester is intervened-on, the acid level is independent of both ester and perfume, is hardly trivial, especially if participants are also focusing on the relation between ester and perfume. Thus, there is a relatively high risk that people mistakenly endorse a link from ester to acid on the basis of a few trials in which the two variables appear to covary. In set-ups with no spontaneous base rate (see general discussion for examples), people are unlikely to be misled in this way.

Finally, as in Experiments 1 and 2, intercue links were underweighted, both in terms of direct endorsement of causal links and a derived measure of contingency. This suggests that another contributor to poor task performance (across all three experiments) might be a prior tendency for people to treat putative causes as independent.

In sum, the temporal cue hypothesis is the only account that can explain the data from all of the experiments conducted thus far. Our final experiment was focused on this hypothesis.

Experiment 4

The temporal order of experienced events is often a reliable cue for causal structure because effects never precede causes. This suggests a useful heuristic: Use experiential order as a proxy for real-world order. That is, if you perceive or produce an event, infer that any subsequent correlated changes are effects of that cause (or effects of an earlier common cause). Such a temporal cue is especially useful in intervention because our actions precede their effects (both in experience and in the world). It can also be used in observation, although it will not be as reliable an indicator because sometimes we receive information about causes after we receive information about effects (especially when making diagnoses).

One critical difference between the observation and intervention conditions in Experiment 1 and 2 was the presence of a temporal priority cue in the latter but not the former. An intervener first selected an intervention (by clicking on a button), and then viewed the values taken by the other variables (by clicking on another button). This contrasts with the observational case, in which all values were displayed simultaneously. Interveners could exploit this cue by assuming that any changes in the values of variables following an intervention were effects of it. By using this temporal cue, which is built into the nature of an intervention, they could more readily identify the effects of a potential cause. This contrasts with the observation condition in which no temporal cues to distinguish causes from effects were presented.

We also explained the results of Experiment 3 in terms of this temporal cue hypothesis. Only the forced observation condition lacked a temporal cue, and in this condition participants performed worst. The other three conditions all involved some temporal delay between displays of information, and model construction performance across them was comparable and better than the forced observation condition.

The hypothesis to be explored in this experiment is that the advantage of intervention over observation is driven by the presence of a temporal cue in the former. To test this, we created four new conditions by crossing the type of learning (observation vs. intervention) with temporal delay (consistent vs. inconsistent). Thus, in time-consistent observation, participants received information about variable values in a temporal order consistent with the causal chain model. In particular, there was a temporal delay between the two putative causes (e.g., acid and ester). In contrast, participants in time-inconsistent observation received information about the two putative causes simultaneously. This is inconsistent with a chain but consistent with a common effect model. Similarly, in time-consistent intervention, after participants have made their intervention, they receive information about the variable values in a temporal order consistent with the chain model; in time-inconsistent intervention, this order was instead consistent with a common effect model.[6]

The four conditions are shown in Figure 15. The critical contrast here is that in the time-consistent conditions there is a temporal delay between putative causes, whereas in the time-inconsistent condition there is no delay. The temporal cue hypothesis predicts an advantage for time-consistent conditions but no general advantage for intervention over observation.

Method

Participants and apparatus. Twenty undergraduates from University College London received $7 each for their participation. None had taken part in any of the previous experiments. All participants were tested individually, and the entire experiment was run on a personal computer.

Procedure and materials. The experiment had a mixed design, with type of learning (observation vs. intervention) as a between-subjects factor, and temporal delay (consistent vs. inconsistent) as a within-subject factor. All participants received the same introductory instructions and cover stories (a chemist and a space engineer scenario) as in previous experiments. In addition, they were warned that they would experience slight temporal delays in the receipt of variable values.

Participants in the observation group carried out both a time-consistent and a time-inconsistent observation task (task order counterbalanced). In both tasks, observers were exposed to the same set of probabilistic learning data used in Experiment 3 (see Figure 11 and Table 3) with order randomized. In the time-consistent condition, the order of display of these data was consistent with the causal chain structure. That is, a short temporal delay (1 s) obtained between the receipt of each variable value. For example, in the chemical scenario, the value for acid was displayed first, followed by a 1-s delay, then the value for ester, and then another 1-s delay before the display of the outcome (perfume present or absent). In the time-inconsistent condition, the order of display of information was inconsistent with the causal chain structure but consistent with a common effect model. That is, the values for acid and ester were displayed simultaneously, followed by a 1-s delay before the display of the outcome.

Participants in the intervention group also carried out both a time-consistent and a time-inconsistent task. In both tasks, interveners were able to make interventions on the same causal chain model used in Experiment 3. However, whereas in previous intervention conditions participants viewed the result of an intervention on a button click immediately after they had made that intervention, in these intervention conditions they had their intervention, and the values of the other variables, displayed with 1-s

[6] There are a variety of patterns of temporal delays that could have been used, each consistent with a different kind of causal model. Future research is planned to explore these alternatives, but for the purposes of this article, a pattern of delays consistent with a common effect is sufficient to make the necessary contrast with the chain model.

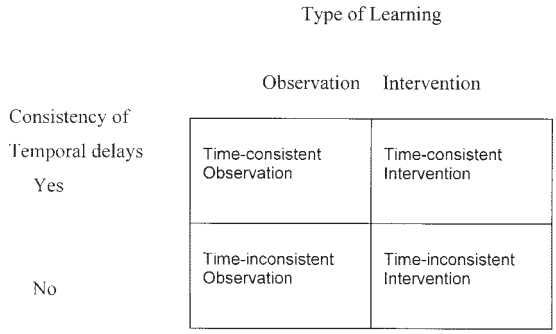

Figure 15. The four conditions in Experiment 4 resulting from crossing type of learning (observation vs. intervention) with temporal delay (consistent vs. inconsistent).

temporal delays. Thus, once the interveners had made their intervention, they viewed the variable values in the same fashion as the observers. In the time-consistent condition, the order was consistent with the causal chain structure, with a 1-s delay between each variable in the chain. In the time-inconsistent condition it was inconsistent, with two of the variables (including the intervened-on variable) being displayed simultaneously, followed by a 1-s delay before the display of the outcome.

The test phase for both groups of participants was identical to Experiment 3. Participants answered a set of conditional probability questions and registered their choice of causal model by clicking on individual links (see Figure 12).

Results

Model selection. As in Experiment 3, a wide range of models were selected in each of the four conditions (the chance choice probability of each model is 0.0625). Table 5 shows the proportions for the seven most common models (these make up over 75% of the total choices). We conducted separate chi-square tests on each condition to see whether the frequency of correct chain models differed from chance. In the two time-consistent conditions, the proportions of participants who constructed the correct chain were significantly greater than chance: time-consistent intervention condition (40%), $\chi^2(1, N = 10) = 45.60$, $p < .001$; time-consistent observation condition (50%), $\chi^2(1, N = 10) = 65.33$, $p < .001$. In contrast, in the time-inconsistent conditions, neither proportion differed from chance: time-inconsistent intervention condition (10%), $\chi^2(1, N = 10) = 0.56$, *ns*; time-inconsistent observation condition (0%), $\chi^2(1, N = 10) = 1.33$, *ns*.

As in the previous experiment, we carried out the same analyses for the observation conditions with the Markov equivalent common cause included as a correct response. Here again the proportion of correct choices was greater than chance for time-consistent observation (60%), $\chi^2(1, N = 10) = 90.25$, $p < .001$, but not for time-inconsistent observation condition (10%), $\chi^2(1, N = 10) = 0.25$, *ns*.

In line with the temporal cue hypothesis, significantly more participants selected the chain model in the consistent conditions (45%) than in the inconsistent ones (5%), $t(19) = 3.56$, $p < .01$. This advantage for the temporally consistent condition was also demonstrated by the differences between time-consistent observation (50%) and time-inconsistent observation (0%), $t(9) = 3.00$, $p < .05$, and between time-consistent intervention (40%) and time-inconsistent intervention (10%), $t(9) = 1.96$, $p < .05$, one-tailed. In further support of the claim that temporal cues help drive the intervention versus observation advantage, time-consistent observation and intervention did not differ, $t(18) = 0.43$, *ns*, nor did time-inconsistent observation and intervention, $t(18) = 1.00$, *ns*.

In contrast to Experiment 3, in both temporally consistent conditions, the modal choice was the correct chain. In the inconsistent observation, the modal choice was the common effect model (30%) and the second most common choice was the single link model (20%). These are comparable with the choice proportions in the observation conditions in Experiments 1–3 that lacked temporal cues. In the inconsistent intervention condition, choices were spread through a variety of models, with a tendency to endorse models with too many links.

Judgments of conditional independence. The correct chain model implies that ester screens-off acid from perfume and thus $P(P|A\&E) = P(P|E)$. The mean ratings for these two probabilities in each of the four conditions are shown in Figure 16. We conducted a mixed ANOVA with probability judgment ($P[P|A\&E]$ vs. $P[P|E]$) and temporal consistency (*yes* vs. *no*) as within-subject factors and type of learning (observation vs. intervention) as a between-subjects factor. The probability being judged had a main effect, $F(1, 18) = 47.06$, $p < .001$, but temporal consistency, $F(1, 18) = 1.69$, *ns*, type of learning, $F(1, 18) = 2.52$, *ns*, and the interactions did not. This parallels the results from Experiment 3 and shows that overall participants' probability judgments were insensitive to screening-off.

Table 5
Proportion of Models Chosen in Each Condition in Experiment 4

Condition	a → e ↙ p	a ← e ↙ p	a → e ↘ ↙ p	a e ↘ ↙ p	a e ↙ p	a ← e ↘ ↙ p	a ⇄ e ↘ ↙ p
	Correct chain	Common Cause 2	Hybrid 1	Common effect	Single link	Hybrid 2	All
Observation time-consistent	0.5	0.1	0.0	0.0	0.1	0.0	0.1
Observation time-inconsistent	0.0	0.1	0.0	0.3	0.2	0.0	0.1
Intervention time-consistent	0.4	0.1	0.1	0.0	0.1	0.0	0.2
Intervention time-inconsistent	0.1	0.1	0.2	0.0	0.0	0.2	0.1

Note. a = acid; e = ester; p = perfume.

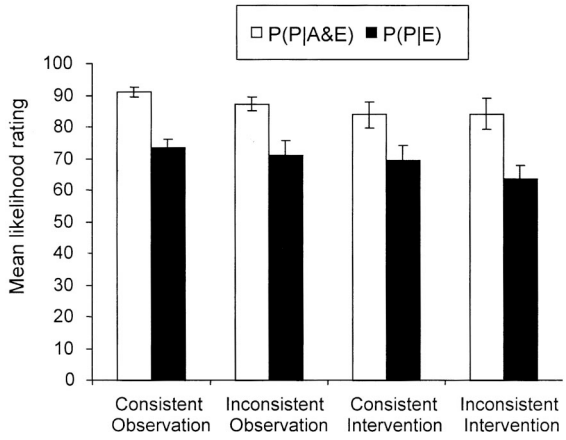

Figure 16. Assessment of screening-off: Mean conditional probability judgments for each condition in Experiment 4.

As in Experiment 3, we separated out the probability judgments of those participants who gave a correct model. These data are displayed in Figure 17. Once again, we grouped time-consistent and time-inconsistent conditions together because there were too few correct responses in the time-inconsistent conditions to make meaningful comparisons across these two conditions. Using these probability judgments, we conducted a mixed ANOVA with probability judgment ($P[P|A\&E]$ vs. $P[P|E]$) as a within-subject factor and type of learning (observation vs. intervention) as a between-subjects factor. In contrast to Experiment 3, there was a main effect of probability judgment, $F(1, 10) = 27.24$, $p < .001$, a marginally significant effect of type of learning, $F(1, 10) = 4.8$, $p = .053$, and a significant interaction between judgment and type of learning, $F(1, 10) = 5.13$, $p < .05$.

This analysis shows that observers who constructed a correct model (six out of seven were in the time-consistent condition) failed to obey screening-off. This is confirmed by the fact that their mean judgments for $P(P|A\&E)$ was significantly greater than for $P(P|E)$, $t(6) = 8.44$, $p < .001$. In contrast, for interveners, there was no significant difference between the two sets of judgments, $t(4) = 1.44$, *ns*, $(1 - \beta = 0.286)$. These findings are reinforced in the individual data, in which none of the six observers who chose a correct model obeyed screening-off, whereas three of the five interveners did. The small sample size, however, gives only weak support to a conclusion of no difference. None of our general conclusions hinge on this particular result.

This discrepancy between observers' model construction and their conformity to screening-off contrasts with the findings in Experiments 1 and 3. It suggests that the presence of consistent temporal delays can promote structure learning even when people are ignorant of the appropriate probabilistic relations. This will be discussed in the next section.

Derived judgments of contingency. As in Experiments 1–3, there was no significant difference between groups; thus, we collapsed across them. Once again the main finding was that the judged contingency between the ester and perfume variables, derived $\triangle P_{EP} = 0.42$, was higher than that between acid and ester, derived $\triangle P_{AE} = 0.32$, $t(39) = 1.87$, $p < .05$, one-tailed. However, this difference was far less than in any of the previous experiments.

Discussion

The main finding was that participants made significantly more correct choices in those conditions in which the temporal display of information was consistent with the causal chain than in those in which it was inconsistent. This supports the hypothesis that the presence of a temporal cue considerably improves the learning of causal structure. Moreover, the lack of overall difference between interveners and observers suggests that the advantage of intervention is largely driven by these temporal cues.

A secondary finding was that none of the observers who selected the correct model gave probability judgments that obeyed the screening-off relations encoded by this model. This contrasts with the close fit between model choice and screening-off for the interveners in the previous experiments, but echoes the partial dissociation between the two noted in Experiment 2 for observers of intervention.

One explanation for this finding is that the temporal cue in the time-consistent observation condition was sufficiently strong; thus, observers could use it to determine the causal structure without recourse to any computations based on probabilistic dependencies. Observers simply assumed that the order of display of information matched the actual causal order and paid less attention to the conditional frequencies of the various events. In contrast, the interveners, because they still had to make interventions on each trial, were more sensitive to these conditional frequencies.

General Discussion

Summary of Results

The first experiment demonstrated an advantage of intervention over observation for the learning of causal structure, both with respect to correct model choices and probability judgments that obeyed screening-off. Subsequent experiments aimed to identify the factors that drove this advantage. Experiment 2 showed that it was not solely due to differences in the distributional information that people were exposed to by finding an advantage for interven-

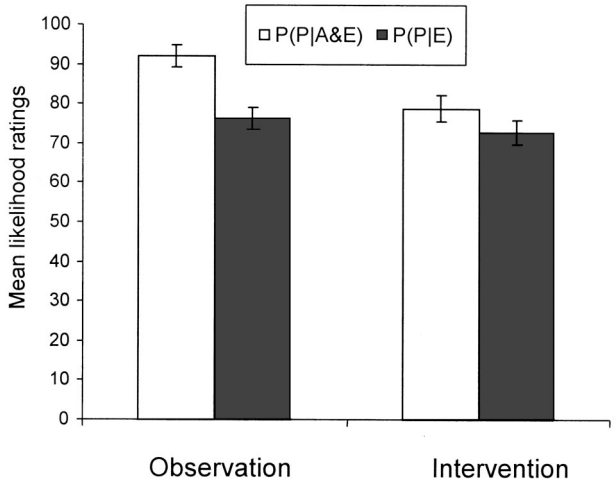

Figure 17. Assessment of screening-off: Mean conditional probability judgments for interveners and observers who chose the correct model in Experiment 4. P = perfume; A = acid; E = ester.

ers while equating the learning distributions. Experiment 3 ruled out two further explanations for the advantage of intervention: one based on interveners being free to make their own selections and thus engage in hypothesis testing, the other based on an increased focus due to the need to decide what intervention to make. Although observers who were able to select their own information did perform better than those who could not, interveners who were forced to carry out prespecified interventions (and thus made no selections or decisions) performed no worse than normal interveners.

The temporal cue hypothesis can account for the model choice data for all these experiments. This is because whenever an intervener makes an intervention (whether freely chosen or prespecified), they experience the putative cause before any of its possible effects. This temporal order can be used as a cue to causal structure. In short, participants could exploit this cue by assuming that any changes in the values of variables following an intervention were effects of it.

A similar argument holds when people are able to select their own observations. Their initial selection precedes their experience of subsequent events, and their inferences about causal structure can be guided (and misguided) by this temporal cue. In the normal observation condition, in which all information about a trial is presented simultaneously, no temporal cues exist, and performance is correspondingly poor.

This hypothesis was directly tested in Experiment 4, and the results suggest that the advantage of intervention is driven by the temporal cue. When information is displayed with temporal delays that mirror the correct causal structure, learning was enhanced. In contrast, when it was displayed with delays that were inconsistent with this structure, learning was impeded. Moreover, this difference held irrespective of whether people intervened or observed.

A second finding in Experiment 4 was that none of those observers who constructed the correct model gave probability judgments that obeyed screening-off. This showed that the veridical temporal delays were sufficient for people to learn causal structure even when their probability judgments did not fit the correct model.

Temporal Cue Heuristic

As we noted in the introduction, interventions differ from observations because they can modify the causal system under investigation by overriding causal links that feed into the intervened-on variable. Our studies suggest that when learning causal structure, people do not represent this situation explicitly but use a heuristic that exploits the temporal precedence between their experience of an intervention and any subsequent changes in the environment. The heuristic tells us to infer that any changes that follow an action are effects of that cause. This is often efficient because in our everyday world, actions necessarily precede their effects. However, it is not foolproof. We are misled when changes subsequent to our actions are effects of a confounding cause—a factor that is either coincidentally related to our actions or a common cause of both our actions and the observed effect. In the case of observation, this heuristic will be a less reliable indicator because the order in which we receive information about causes and effects will not always reflect their real-world order. However, as long as there is some temporal delay between the displays of information, people still seem to employ the heuristic. This is evidenced by the inferences made by observers who make their own selections prior to viewing the values of other variables (the selected observation condition in Experiment 3). Changes subsequent to their selections are encoded as effects of that putative cause, as shown by their tendency to infer either the correct causal chain or its Markov equivalent common cause model.

This is not to deny that temporal order is a good guide to causal order, even in cases of observation. We often experience causally related events in the same order as that in which they occur in the world. This is particularly true when we perceive the operation of a physical system or process. In such cases, temporal order does provide a very stable cue as to causal order. However, in situations in which the receipt of information about events can be delayed differentially, we no longer have the assurance that the order of our experiences matches the true causal order.[7]

In sum, our findings suggest that people use a simple temporally based heuristic: Once they make an intervention or selection, they infer that any subsequent changes are effects of this cause. This is sufficient to explain an advantage of intervention over observation, because an intervention overrides the action of any other causes leading into the intervened-on variable; thus, it can reasonably be held responsible for subsequent changes. In contrast, an observation of one event followed by another event comes with no such guarantee, because it is possible that there is a common cause of both or that the temporal order in which these events are experienced does not match their real-world order. The temporal cue heuristic exploits this difference without requiring that people reason about it explicitly.

Tendency to Treat Causes as Independent

In all four experiments, participants showed a tendency to underweight the relation between putative causes in comparison with the relation between a cause and the outcome. This was evidenced by the derived measures for the judged strengths of these links and by participant's model selections. This tendency was particularly apparent in the pure observation conditions, with a strong bias in favor of the common effect model or a single link model from one cause to the outcome.

A simple explanation for this is that the experimental task (and the content of both scenarios) encouraged participants to focus on predicting the outcome. Consequently, in the learning phase, people were likely to pay more attention to the links from putative causes to the outcome (e.g., the acid → perfume and ester → perfume links in the chemist scenario) and less attention to possible links between these variables (e.g., acid → ester). This tendency was exaggerated when all three variable values were presented simultaneously (normal observation conditions in Experiments 1–3) or when both putative causes were presented together followed by the outcome (time-inconsistent condition in Experiment 4). It was attenuated, however, when there were temporal delays between the putative causes (time-consistent condition in Experiment 4).

[7] It is also possible to receive information about an intervention after having observed its effect. However, this typically only happens for observers of intervention rather than interveners themselves.

This is not to claim that this prediction bias is merely an experimental artifact. In everyday life, we are often focused on the prediction of a target outcome on the basis of multiple cues or causes, and a reasonable default strategy is to assume that these cues or causes are independent. In a similar vein, many statistical procedures default to the assumption that predictor variables are independent (e.g., multiple linear regression).

Relation Between Model Choice and Screening-Off

What do our experiments tell us about the relation between people's choices of model and their probability judgments? Recent analytic work (Pearl, 2000; Spirtes et al., 1993) has provided us with a theoretical account of how the two should be related. Each causal structure imposes constraints on the probabilistic dependencies it can generate, and, likewise, certain sets of probabilistic dependencies imply particular sets of causal models. For example, the chain structure ($A \rightarrow B \rightarrow C$) used in our experiments implies that A, B, and C are (unconditionally) dependent and that A and C are independent conditional on B (screening-off). Therefore, if one knows that these dependencies hold, one can infer that the generating structure is either the $A \rightarrow B \rightarrow C$ chain or its Markov equivalent common cause $A \leftarrow B \rightarrow C$. One question our experiments raise is whether people actually make an inference from dependencies to causal structure.

Apparent evidence in favor of this is the close fit between model choices and conformity to screening-off for interveners who chose the correct model in Experiments 1–3. However, this is equally consistent with the possibility that conformity to screening-off is determined by people's causal models rather than vice versa. Moreover, there are several discrepancies between model choices and screening-off (e.g., for the observers of intervention in Experiment 2 and for observers in the temporal delay conditions in Experiment 4). Finally, under normal observation conditions, participants uniformly failed to infer the correct causal models, even though the requisite information was present in the data they experienced. This raises questions about their ability to infer structure on the basis of covariational information alone.

Constraint-Based Versus Causal-Model Approaches to Structure Learning

Overall, our results suggest that sensitivity to the probabilistic dependencies in the learning data is neither necessary nor sufficient for inferring causal structure. This has implications for the debate between the different computational approaches to causal induction. For one, it suggests that people do not use purely constraint-based methods for inferring causal structure. As previously noted, such methods require that people build up candidate models from the conditional independencies in the data. However, participants in our experiments failed to notice these conditional independencies unless they got the model right; they often chose models (e.g., common effect models) that were incompatible with the data and they underestimated objective contingencies that did not underlie assumed causal links within their chosen models.

Taken together, these findings are more supportive of the causal-model approach (Waldmann, 1996), in which prior assumptions or knowledge guide the induction process. In particular, assumptions about temporal priority seem to play a crucial role in our experiments. Participants appear to generate candidate causal structures based on the available temporal cues and encode the learning data in the light of these models. This explains a variety of the current findings: the underestimation of relations that are not marked by a temporal cue (e.g., the link between the two cue variables in observation conditions), the match between screening-off judgments and model choices when the correct model is inferred, the under-representation of statistical patterns in the data that are inconsistent with the chosen model.

It also explains how people can infer the correct model on the basis of temporal cues, even though their judgments do not respect screening-off (e.g., in Experiment 4) and infer the incorrect model when their judgments do respect screening-off (e.g., observation of intervention condition in Experiment 2). Finally, it reinforces our explanation of the bias toward a common effect model in the normal observation conditions. The absence of any temporal delay in the presentation of two of the variables (e.g., acid and ester) reduces the chances of people detecting that the two are causally linked.

In sum, the overall pattern of results across our experiments is supportive of the causal-model approach to structure learning and offers little support for a purely constraint-based approach. The causal-model approach places fewer constraints on the learning process, however, and therefore is harder to disconfirm. Nevertheless, our data suggest that prior assumptions about the relation between temporal and causal priority appear to direct learning in this paradigm.[8]

Relevance of Temporal Cues in Causal Learning

The idea that temporal cues play a crucial role in learning causal structure is perhaps not too surprising. However, its importance has not been recognized in much contemporary research due to the focus on how people estimate the causal strength of presorted causes and effects and the inattention to the potential differences between intervention and observation. There is, however, some recent research that underlines the important role that our knowledge of temporal delays plays in causal learning. Hagmayer and Waldmann (2002) demonstrated that prior assumptions about the temporal delays between events determine which events people pair together as cause and effect and what they select as a possible cause. Similarly, Buehner and May (2002) showed that people can use prior expectancies about temporal delays to modulate their ratings of the strength of a causal relation. Relatedly, Gallistel and Gibbon (2000) argued that the core process in animal conditioning is the learning of temporal intervals and rates of event occurrence. Although yet to be applied directly to human learning, it too highlights the idea that temporal aspects of experience contribute greatly to successful learning.

This work suggests that people's use of a temporal cue heuristic for inferring causal structure might be modulated by their prior knowledge about the kinds of delays they should expect. For example, if you take a vitamin supplement to improve the condition of your skin, you expect it to work in a few weeks rather than a few seconds. In contrast, administering a fake tan lotion can be expected to work in seconds or minutes rather than days. Whether

[8] This section is based on suggestions made by Michael Waldmann.

the temporal cue heuristic is restricted to events relatively close in time or whether it can be used with longer delays is an open question. We suspect that more automatic learning will be restricted to a limited time window, whereas more reflective learning can stretch across longer time delays.

Other Research on the Difference Between Observational and Interventional Learning

Although we have argued that the special information provided by intervention is not what drives the advantage in our studies, it is possible that in different kinds of learning tasks people can make more explicit use of this information. For example, it is possible that in a simpler two-variable task, people would be able to infer causal structure on the basis of interventions alone, even when the temporal order of receipt of information is arranged to conflict with the causal order.[9]

The complexity of the task may be an important factor. Recent work by Steyvers et al. (2003), Sobel (2003), and Lagnado and Sloman (2003) showed that when the task of detecting correlations is relatively simple, people are better able to distinguish three-variable causal structures by means of appropriate interventions.

Steyvers et al. (2003) used a task in which participants had to induce the communication network between a group of three alien mind readers. The use of categorical variables with many possible states (words) allowed correlations between different mind readers to be established in a few trials. Further, the tracking of the effects of an intervention was facilitated by using a distinctive word that was "implanted" in a selected alien's head. People were tested on a variety of three-variable structures and even with the choice of just one intervention (which was then repeated on numerous trials), participants performed substantially above chance (30% to 35% correct responses) and better than when given just observational data. The presentation of data (e.g., the words in each of the alien heads on each trial) was simultaneous, although of necessity, an intervention was selected before its effects could be viewed.

The authors also modeled people's choice of intervention within a rational framework (Tong & Koller, 2001) and found that across the various different structures (common effect, common cause, chain, single-link) a high proportion of these interventions were close to optimal with regard to maximizing information about causal structure. The exception here was the chain structure, in which participants preferred to intervene on the first (source) node of a chain rather than on the more diagnostic intermediate node. However, it is not clear from the averaged results whether this bias led to a correspondingly poor discrimination of chain structures. Overall, these findings are explicable in terms of the modification-based information unique to interventions; however, they are also consistent with the operation of a temporal heuristic.

The studies by Sobel (2003) involved networks of colored lights with color-sensitive sensors. Participants either observed trial-based combinations of these lights (on or off) or made interventions by activating one light at a time and observing its effects. Performance across a set of five three-variable models was relatively high (66% in intervention), although participants were unable to distinguish a causal chain from a structure with an additional link. Here again, one possible conclusion is that people use modification-based information to infer causal structure. However, the patterns of results are also explicable in terms of temporal cues. Lagnado and Sloman (2003) used a real-time paradigm in which people either observed or manipulated a network of on-screen sliders. The presentation format greatly facilitated the detection of correlations between components and led to a very high proportion of correct model selections (over 90%). Moreover, with chains, people were able to make the additional intervention required to uniquely identify the true structure.

One important factor that differentiates these studies from the current experiments is that their learning environments prevented interveners from being misled by spurious correlations between their interventions and subsequent changes in other variables. This factor is likely to have boosted interveners' ability to learn causal structure. It guarantees that after intervening on a particular variable, any subsequent changes in other variables are indeed effects of that cause. In contrast, in our current experiments, participants were readily misled, because there was a relatively high probability (0.5) that a spurious change would occur after an intervention on the intermediate variable of a chain.

What all of these studies suggest is that if the initial task of detecting correlations between variables is facilitated, people are better able to use interventions to infer causal structure. Although the use of modification-based information cannot be ruled out, this advantage is sufficiently explained by the temporal cue associated with intervention.

Relevance to Models of Causal Learning

The predominant psychological models of causal learning are either rule-based (e.g., Allan, 1993; Cheng, 1997) or associative (e.g., Shanks & Dickinson, 1987; Shanks, 1995; Van Hamme & Wasserman, 1994). These have typically been applied to the estimation of causal strength given a known causal structure, usually a common effect model. Although neither have been directly applied to the kind of structure learning examined in this article, it is possible to anticipate how the frameworks might be applied to accommodate our results.

Rule-based models assume that people encode appropriate event frequencies during the learning phase and then compute causal strengths according to some integrating rule. By extension we could assume that they explain structure learning in a similar fashion by computations on the basis of the encoded frequency data. Our findings, however, indicate that people do not seem able to infer the correct causal structure on the basis of atemporal data alone, as evidenced by their poor performance in pure observation conditions. Furthermore, such an account gives no explanation for the advantage of intervention over observation. This does not mean that a rule-based account could not be augmented to deal with these factors; however, a satisfactory theory must explain how the temporal cues provided by people's interventions or selections enhance their ability to learn causal structure.

Associative models assume that people engage in the predictive learning of effects from causes (or vice versa) and that they base their causal judgments on these learned associations. The bias that we found in favor of a common effect model (and in later experiments a single link model) would be explicable on an associative account under the assumption that observers are treating the two

[9] This example was suggested by Michael Waldmann.

putative causes as independent predictors of the outcome. How an associative model would deal with the distinction between observation and intervention is less clear. The predominant learning rule, the Rescorla-Wagner rule (1972), includes parameters for the salience of both cues and outcomes; thus, it is possible that the distinction between observational and interventional learning could be marked by differences in salience. However, this would be entirely ad hoc. A more plausible approach is to note that the importance of temporal ordering is in some sense built into the predictive learning rule, in which an expectation is generated on the basis of the observed cues and then corrected in the light of the actual outcome. The distinction between observation and intervention need not be marked by the associative learning mechanism per se but by the fact that, in the case of intervention, associative bonds are more likely to be built up from genuine causes to genuine effects. Once again, this is a natural consequence of the fact that actions precede their effects.

Although this kind of approach may suffice for situations in which people learn in a more intuitive and nonreflective manner, it may not scale up to cases in which people deliberate about possible causal structures and carry out appropriate tests (e.g., quasiscientific experimentation). A related problem for an associative account is to give a plausible account of how people learn more complex causal structures—how do people stitch together simple associations to create models with multiple connections?

Conclusions

A critical difference between intervention and observation is that in the former one can modify the causal system under investigation. This difference is formalized within the Causal Bayes net framework (Pearl, 2000; Spirtes et al., 1993), and Sloman and Lagnado (2002, 2003) showed that when people reason on the basis of known causal structure, they are able to represent interventions distinctly from observations. In contrast, our current experiments suggest that when inducing causal structure, people do not represent the difference between intervention and observation explicitly but use a temporal heuristic in both learning contexts. The advantage for intervention derives from the fact that the temporal cue is a more stable indicator of actual temporal order when one is intervening on a system rather than when one is passively observing it. In short, actions must always precede their effects, whereas the temporal order in which we receive information about events or variable values need not always reflect the order in which they actually take place in the world.

We do not suggest that this temporal heuristic excludes the possibility of other routes for interventional learning. In particular, both modification and selection-based information are likely to play a significant role in certain induction tasks. However, interventions can aid causal structure learning without an individual having to engage in sophisticated computations. Interventions may improve learning because they modify the system under investigation rather than the representation of the system.

This conclusion implies that there is a discontinuity between the unreflective causal learning investigated in this article and the more deliberate hypothesis testing characteristic of experimental sciences. The latter involves explicit representation of the informational consequences of experimental manipulation, whereas the former relies on more primitive associative processes (cf. Evans & Over, 1996; Sloman, 1996). This is not to deny the crucial role that intervention plays in the discovery of causal structure but to suggest that our cognitive mechanisms sometimes exploit rather than understand the difference between experiment and observation.

References

Ahn, W., & Dennis, M. (2000). Induction of causal chains. In *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 19–24). Mahwah, NJ: Erlbaum.

Allan, L. G. (1980). A note on measurements of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society, 15,* 147–149.

Allan, L. G. (1993). Human contingency judgments: Rule-based or associative? *Psychological Bulletin, 114,* 435–448.

Bacon, F. (1620). *Novum Organum*. Chicago: Open Court.

Buehner, M. J., & May, J. (2002). Knowledge mediates the timeframe of covariation assessment in human causal induction. *Thinking and Reasoning, 8,* 269–295.

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

Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove, England: Psychology Press.

Gallistel, C. R., & Gibbon, J. (2000). Time, rate, and conditioning. *Psychological Review, 107,* 289–344.

Glymour, C. (2001). *The mind's arrows*. Cambridge, MA: MIT Press.

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,* 1–31.

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge, England: Cambridge University Press.

Hagmayer, Y., & Waldmann, M. R. (2002). How temporal assumptions influence causal judgments. *Memory & Cognition, 30,* 1128–1137.

Hammond, K. R. (1996). *Human judgment and social policy*. New York: Oxford University Press.

Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world*. Thousand Oaks, CA: Sage.

Heckerman, D., Meek, C., & Cooper, G. (1999). A Bayesian approach to causal discovery. In C. Glymour & G. Cooper (Eds.), *Computation, causation, and discovery* (pp. 143–167). Cambridge, MA: MIT Press.

Hume, D. (1748). *An enquiry concerning human understanding*. Oxford, England: Clarendon.

Lagnado D. A., & Sloman, S. A. (2003). *Dynamic learning of causal structure*. Manuscript in preparation.

Mackintosh, N. J., & Dickinson, A. (1979). Instrumental (Type II) conditioning. In A. Dickinson & R. A. Boakes (Eds.), *Mechanisms of learning and motivation* (pp. 143–169). Hillsdale, NJ: Erlbaum.

Mill, J. S. (1950). *Philosophy of scientific method*. New York: Hafner. (Original work published 1843)

Pearl, J. (2000). *Causality*. Cambridge, England: Cambridge University Press.

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 theory and research* (pp. 64–99). New York: Appleton-Century-Crofts.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shanks, D. R. (1995). Is human learning rational? *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 48,* 257–279.

Shanks, D. R. (in press). Judging covariation and causation. In D. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making*. Oxford, England: Blackwell.

Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgment. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 21, pp. 229–261). San Diego, CA: Academic Press.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119,* 3–22.

Sloman, S. A., & Lagnado, D. A. (2002). Counterfactual undoing in deterministic causal reasoning. In W. Gray & C. D. Schunn (Eds.), *Proceedings of the twenty-fourth annual conference of the cognitive science society* (pp. 828–833). Mahwah, NJ: Erlbaum.

Sloman, S. A., & Lagnado, D. A. (2003). *Do we "do"?* Manuscript submitted for publication.

Sloman, S. A., & Lagnado, D. A. (2004). Causal invariance in reasoning and learning. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 44, pp. 287–325). San Diego: Elsevier Science.

Sobel, D. (2003). *Watch it, do it, or watch it done*. Manuscript submitted for publication.

Spirtes, P., Glymour, C., & Schienes, R. (1993). *Causation, prediction, and search*. New York: Springer-Verlag.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & 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. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems 13* (pp. 59–65). Cambridge, MA: MIT Press.

Tong, S., & Koller, D. (2001). Active learning for structure in Bayesian networks. *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence,* 863–869.

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. (1996). Knowledge-based causal induction. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation* (Vol. 34, pp. 47–88). San Diego, CA: Academic Press.

Waldmann, M. R. (2001). Predictive versus diagnostic causal learning: Evidence from an overshadowing paradigm. *Psychonomic Bulletin & Review, 8,* 600–608.

Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition, 82,* 27–58.

Received August 5, 2003
Revision received December 3, 2003
Accepted December 4, 2003 ■