Interventions do not solely benefit causal learning: Being told what to do results in worse learning than doing it yourself

David M. Sobel (David_Sobel_1@brown.edu)
Department of Cognitive and Linguistic Sciences, Box 1978, Brown University
Providence, RI, 02912 USA

Tamar Kushnir (Tkushnir@socrates.berkeley.edu)
Department of Psychology, Tolman Hall, University of California
Berkeley, CA 94720 USA

Abstract

Previous research has demonstrated that causal learning is facilitated by observing interventions on a causal system (e.g., Lagnado & Sloman, 2002). Does the origin of these interventions influence learning? Sobel (2003) demonstrated that causal learning was facilitated when learners observed the results of their own interventions as opposed the results of another's interventions, even though the data learners observed were identical. Learners in the former condition were able to test various causal hypotheses, while learners in the latter condition were less able to do so. The present experiment followed up on these findings by comparing causal learning based on observing the results of a learner's own interventions with causal learning based on observing data from a set of interventions a learner is forced to make. Although learners observed the same interventions and subsequent data, learning was better when participants observed the results of their own interventions. These findings are discussed in relation to various computational models of causal learning.

Introduction

Causal knowledge is important for everyday interaction in the world. A recent pursuit in cognitive science has been to describe how human beings learn and represent causal knowledge. Researchers in computer science have examined causal graphical models as a way of representing causal structure. Recently, psychologists have adopted this formalism as a way of representing causal relations in a variety of domains (Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks, in press; Lagnado & Sloman, 2002; Rehder & Hastie, 2001; Tenenbaum & Griffiths, 2002).

One issue in this discussion is whether a particular algorithm or class of algorithms best instantiates human causal learning. Glymour (2001) proposes that constraint-based models offer the best account of human causal learning (see also Gopnik et al., in press). These models posit rules for learning causal structure based on observing patterns of dependence and independence among a set of events. In contrast, Tenenbaum and colleagues (Tenenbaum & Griffiths, 2002; Steyvers et al., in press) propose that human causal structure learning is better instantiated by Bayesian algorithms. Using these algorithms, learners assign probabilities to a constrained set of hypotheses. They update those probabilities based on the observed data through an application of Bayes' rule. The resulting posterior probabilities on the hypotheses represent how likely each is the causal structure that generated the observed data.

These algorithms make different predictions about the role of interventions in causal learning. Constraint-based algorithms treat interventions as special conditional probabilities, which enable learners to distinguish between otherwise equivalent causal graphs (Pearl, 2000). Constraint-based algorithms ignore the source of the interventions: only knowledge of conditional independence and dependence is critical for learning.

In contrast, Bayesian algorithms require that learners have particular hypotheses in mind, and given the observed data, update the probability that each hypothesis reflects the actual causal structure. If a learner observes interventions that are relevant to those hypotheses, learning will be facilitated. However, if a learner observes interventions that are not relevant to those hypotheses, then those data will be less beneficial, even if those data contain the critical conditional independence and dependence information necessary to discern a unique causal structure.

An (albeit simple) example might help. Suppose you wake up one morning with the flu, and have two symptoms: coughing and dry-mouth. Many different causal models follow from these observations. For example, the flu could cause the dry-mouth, which in turn could cause coughing. Alternatively, the flu could independently cause coughing and dry-mouth: coughing and dry-mouth could be unrelated, but dependent given that you have the flu. Figure 1 depicts these two potential causal models.

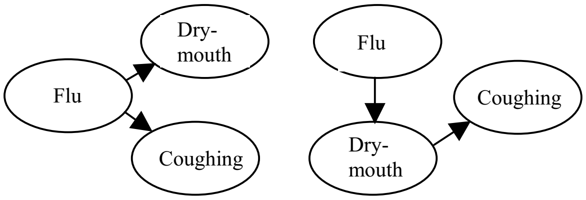

Figure 1: Two potential causal models

An intervention could distinguish between these two hypotheses: drinking a glass of water would eliminate the dry-mouth, but have no effect on the flu. Learners would observe the probability of coughing given the presence of the flu, but the absence of dry-mouth. If coughing persisted, then the left-hand model of Figure 1 would be the more likely causal structure. If coughing were eliminated, then the right-hand model in Figure 1 would be more likely.

However, this "drink water" intervention is only effective at discriminating among certain causal models, such as the two shown in Figure 1. For example, another hypothesis is that dry-mouth and flu both cause coughing, and are independent from each other (perhaps because you are also dehydrated). The "drink water" intervention does not discriminate between this common effect model and the common cause model depicted on the left side of Figure 1. Given that you have the flu, both models predict that coughing would persist if dry-mouth were eliminated.

What this example illustrates is that a learner's initial hypotheses matter. This is not a novel idea in cognitive science: many researchers propose that adult learners have a variety of hypothesis-testing strategies, which may lead them to proper or improper conclusions (see e.g., Klayman & Ha, 1987; Wason, 1968). Is this the case in the domain of causal learning? If learners use constraint-based algorithms to build a causal model from observed data, then the source of those data should not matter. However, if learners are actively testing hypotheses, then the source of interventions might make a difference in causal learning.

This hypothesis relies on two assumptions about causal learning. The first assumption is that learners benefit from interventions over simply observing data. Schulz (2001) investigated whether adults and children could make causal inferences based on interventions. In one experiment, she presented participants with two creatures (A and B) that moved together simultaneously, and told them that one was the "boss", and made the other creature move. From this information alone, the directionality of the causal relation between the two events was indeterminate: neither temporal priority nor contingency information allowed the participants to determine the nature of the causal relation between the two creatures. Participants were then shown that there was an intervention (i.e., a button), designed for one of the creatures (B), which made it move. The button was pressed and only creature B moved. After observing this intervention, both adults and 4-year-olds consistently claimed that the other creature (A) was the boss.

This valid conclusion follows from one of Pearl's (2000) algorithms for learning causal structure through interventions. Learners have observed a critical conditional independence: $p(A \mid B) > p(A \mid do(B)) = 0$. Thus, these data are consistent with the predictions of constraint-based model. However, Schulz's (2001) experiments are also consistent with a Bayesian model. The cover story of the experiment specified that one of two hypotheses were correct; the intervention produced data that distinguished between them.

The second assumption of this project is that when learners are able to generate their own interventions, they choose interventions that enable them to test their hypotheses. Steyvers et al. (in press) provided a critical piece of evidence for this assumption. They asked participants to learn a causal structure from observing data. After participants generated a set of structures that reflected this learning, Steyvers et al. (in press) allowed learners to observe the results of one intervention on that structure. Learners did not choose their intervention randomly. The majority of learners chose the intervention that provided them with the information necessary to distinguish among the models they chose previously. Steyvers et al. (in press) presented a Bayesian model with an "active learning" component (e.g., Murphy, 2001), which provided the best model of these data.

However, neither of these experiments directly tested a situation that compared learning from observing the results of one's own interventions with the results of another's interventions. Sobel (2003) presented participants with a novel causal learning task: participants were told that "Dr. Science" had wired up a set of colored lights and sensors. The sensors were color sensitive and made the light they were connected to activate. Thus, if a red sensor was connected to a white light, then the white light would activate whenever the red light activated. Since this happened at the speed of light, learners just saw the white and red lights activating together. Depending on the experiment, the sensors could have deterministic or probabilistic relationships. Participants were asked to learn the causal structure among four colored lights (i.e., whether each light had sensors on it, and if so, of which color). One group of learners was allowed to intervene on those lights themselves. Another group observed another person make the same interventions (with the same results). Over several experiments, although learners in both groups observed identical interventions and data, learning was superior when participants observed the results of their own interventions.

The goal of this investigation is to examine two concerns with Sobel's (2003) experiments. In these experiments, learners actively manipulated the learning environment; observers, in contrast, did not. The first concern is that it is possible that learners in the intervention condition were simply paying more attention than learners in the observation of intervention condition. Sobel (2003, Experiment 4) attempted to control for this by presenting learners with the ability to select cases in which they observed a particular light activating. In this condition, learners actively manipulated the learning environment, but had no information about interventions (and hence, little information about conditional independence and dependence). Learning was quite poor in this situation.

However, a better manipulation would be to allow learners to manipulate a learning environment, but disable them from testing their own hypotheses. To do this, we created a fixed intervention condition. In this condition, the learning environment was identical to an intervention condition, but learners were forced to make a particular set of interventions. This way, learners generated interventions, but those interventions were based on another learner.

This allows us to examine a second concern with Sobel's (2003) experiments, based on literature in educational psychology. Several researchers suggest that kinesthetic learning – the ability to act on the environment as opposed to simply observing – may benefit learners even in non-kinesthetic domains, such as language (e.g., Furuhata, 1999). Such a benefit may also translate to the domain of causal structure learning. The results of the fixed intervention condition allows us to examine whether observing the results of interventions generated by a learner, independent of their own hypotheses impairs causal structure learning compared with learners who observe the results of their own interventions.

Experiment

In this experiment, we presented learners with four causal structure learning problems using the same paradigm as Sobel (2003). Participants were introduced to Dr. Science, and told that he had wired together colored lights and sensors. Participants were asked to learn the causal structure among four colored lights (i.e., whether each light had sensors on it, and if so, of which colors). Participants were either allowed to intervene on the data freely (Intervention Condition), or were instructed to make specific interventions (Fixed Condition). In the fixed condition, these interventions and the subsequent data were yoked to a learner in the intervention condition. Thus, learners observed identical data across the two conditions.

Participants in the intervention condition were allowed to turn on lights, as well as temporarily remove any one light from the causal structure (by placing a bucket over it, which covers it and its sensors). Sobel (2003, Experiment 5) found that this additional type of intervention greatly benefited learning. In particular, this enabled learners to observe conditional independence relations as well as conditional dependence relations – the exact data necessary to learn a causal structure given a constraint-based algorithm. As long as the conditional independence relations are present, these algorithms would predict no difference between these two conditions. However, if hypothesis testing is important to learners, then when they observe the results of their own interventions, they do so based on their own causal hypotheses. Learners in the fixed intervention condition, in contrast, observe interventions and data based on another's hypotheses, which might not match there own, and which might cause impaired learning.

Method

Participants: Forty-eight undergraduates were recruited from an urban area university. Approximately equal numbers of men and women participated in the experiment. Participants were paid $7 for their participation.

Materials: Participants were tested on a Dell Dimension 8100 desktop computer with a 19" monitor.

Procedure: Following Sobel (2003, Experiment 5), all participants were seated at the computer and given the following instructions:

> "In Dr. Science's laboratory, he has created a number of games. Each game has four lights, colored red, white, blue, and yellow. Each light also has zero, one, or many sensors. Some sensors are sensitive to red light, others to blue light, others to white light, and others to yellow light and will activate the light that it is connected to. For example, if the red light is connected to a yellow sensor, then whenever the yellow light activates, the red light will also activate. But, because this happens at the speed of light, all you will see is the red and yellow lights activating together. It is also possible that a light has no sensors attached to it, and therefore is not activated by any other light. Dr. Science is pretty absentminded, so he is not careful about how he wires the lights together. Sometimes the sensors do not always work perfectly, so they won't always turn the lights they are connected to on.
>
> For each game, you have to figure out how the lights and sensors are wired up. To help you, Dr. Science is going to let you turn on each of the lights. So, you will see a set of buttons that turn on each light.
>
> In addition, Dr. Science has given you a black bucket. You can put the bucket over any one of the lights (and the sensors it is connected to). Putting the bucket over the light covers both it and its sensors. If the bucket is over the light, then the light will not activate because the other lights cannot reach its sensors."

Participants were divided into two groups. In the *intervention* group, participants were told that they would be able to turn on any of the four lights as well as place the bucket over any of the lights as many times as they wanted. These participants were shown a computer screen with nine buttons. Four of the buttons activated the four different colored lights. If one of these buttons was pressed, then that light appeared on the screen for 0.5s. In addition, any effect of that light (and likewise any of its effects) also appeared at the same time on the screen. The probability that any effect occurred given that its cause occurred was 0.8. Thus, turning on a light did not always cause its effects, which learners can attribute to Dr. Science carelessness in how the lights and sensors were wired together.

The other five buttons moved the bucket – either over one of the four lights or off all of the lights. Pressing one of these buttons would result in the bucket moving to the appropriate place. The bucket appeared on the screen as a black box.

Participants in this condition were told to intervene as much as they wanted in order to learn how the lights and sensors were wired together. However, they were required to turn on the four lights a minimum of 25 times. Their interventions and the subsequent data were recorded. When they indicated they were finished, they were asked the test questions described below.

Participants in the *fixed intervention* condition were shown the same nine buttons. However, Dr. Science gave them instructions to press particular buttons in a particular order. Participants were not allowed to intervene on their own accord, but could only generated interventions that turned the lights on and moved the bucket based on what Dr. Science told them. These instructions appeared on the screen one at a time. If participants pressed an incorrect button, then they received an error message.

Importantly, each sequence of interventions given to a participant in the *fixed intervention* condition was yoked to a participant in the *intervention* condition. Thus, the first participant in fixed intervention condition was told to make identical interventions (and observed identical results) as the first participant in the intervention condition.

Before learning the four models, participants in both conditions were given a training session with the bucket. They were shown three lights (colored green, purple, and gray), and three buttons that each activated one of the lights. They were told that Dr. Science had wired these lights and sensors perfectly. They were first told to press the button that activated the gray light, and were shown that only the gray light activated. They were told that from this they could conclude that there was not a gray sensor on the other two lights. One could conclude this since if there had been a sensor on either of those lights, those lights would have activated.

Then they were told to activate the green light and observed that green and gray activated together. From this, they were told to conclude that there was a green sensor on the gray light. Then, they were then told to activate the purple light, and shown that all three lights activated. From this, they were told that they could conclude that there was a purple sensor on the green light. However, whether there was a purple sensor on the gray light was uncertain, since they already knew that there was a green sensor on the gray light and that the green light activated.

They were then told that using the bucket could help. In particular, if the purple light was activated with the bucket on the green light, then they could observe whether the purple light activated the gray light in the absence of the green light. Participants were told to make this intervention and shown that under this circumstance only the purple light activated. Thus, participants observed the relevant conditional independence between purple and gray given the absence of green. The probability that the gray light activated given the purple light and not the green light was zero. There must not be a purple sensor on the gray light. Thus, there was only a purple sensor on the green light and a green sensor on the gray light.

After this training, participants were asked to learn four different causal models, a chain, a chain with a link from the root light to the leaf light (A->D), a common effect and chain and a diamond model. These are shown in Figure 2. The color of the lights was randomly determined for each model. The causal connections were always probabilistic. Thus, if a cause occurred, the probability that its effect occurred was 0.8. Importantly, lights never occurred spontaneously: the probability of an effect given the absence of a cause was zero. The models were presented in one of four quasi-random orders. In both conditions, participants were allowed to take notes.

After participants indicated that they thought they knew how the lights and sensors were wired together (in the intervention condition) or were shown the same intervention sequence (in the fixed intervention condition), participants were asked two sets of questions. First, participants were asked a set of *causal structure* questions – whether each light had a sensor of each other color attached to it (e.g., was there a blue sensor on the red light?). Participants first answered the yes/no question, then rated their level of confidence in their answer on a scale of 0-100, with zero being a total guess, and 100 being fully confident. There were twelve of these questions (one for each possible combination), asked in a random order.

Participants were also asked a set of *conditional probability* questions, in which they were asked to rate the likelihood of a particular light activating given that they had turned on another light. These questions were phrased as follows: "suppose you turned on the X light and the bucket wasn't on any of the lights, what is the probability that the Y light (among possible others) would activate". Participants were asked the twelve exhaustive pair-wise combinations, plus four questions that asked about the conditional probability of that light activating by itself (e.g., what is the probability that only the X light comes on if you turn on the X light). The order of these question sets was counterbalanced across participants.

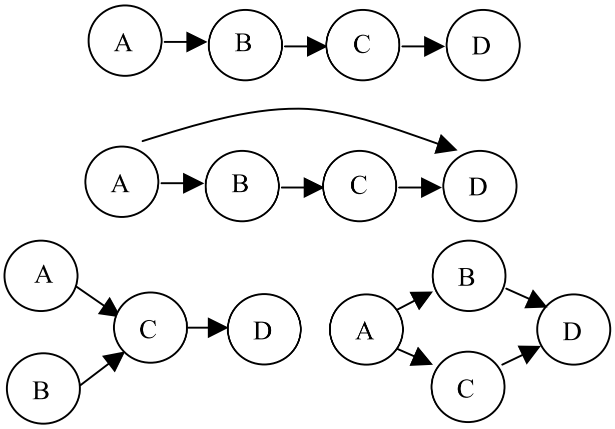

Figure 2: The four causal structures that participants were asked to recover. The colors were randomly assigned for each participant

Results

For each model, participants received a score of one for each causal structure question they answered correctly, and a score of zero for each question they answered incorrectly. These scores were summed to reflect an overall accuracy score for each model (maximum: 12). These scores are shown in Table 1. Preliminary analyses revealed no order effects: neither the question order nor the order the models were presented in influenced responses.

A 4 (model) x 2 (condition) mixed Analysis of Variance was performed on participants' accuracy scores. Model was a within-subject factor; condition was a between-subject factor. This analysis revealed a main effect of condition: overall, learners were more accurate in the intervention condition than the fixed intervention condition: $F(1, 46) = 8.74$, $p < .01$. A main effect of model was not found. No significant interactions were found. Simple effect analysis revealed that this difference was consistent for each of the four models: $t(46) = 2.07$, 2.01, 3.24, and 2.20 for the chain, chain with A->D link, common effect and chain, and diamond models respectively, all p-values $\leq .05$.

Table 1: Accuracy scores on causal structure questions for each model and condition (Maximum = 12). Standard deviation shown in parentheses

Model	Intervention Condition	Fixed Intervention Condition
Chain	10.88 (1.44)	9.88 (1.87)
Chain with A->D link	11.08 (1.10)	9.96 (2.51)
Common Effect and Chain	11.33 (1.47)	9.79 (1.82)
Diamond	10.96 (1.08)	9.88 (2.15)

Table 2: Deviance scores on conditional probability questions for each model across the conditions. Standard deviation shown in parentheses

Model	Intervention Condition	Fixed Intervention Condition
Chain	12.66 (10.37)	14.44 (10.59)
Chain with A->D link	17.23 (11.67)	25.23 (16.19)
Common Effect and Chain	12.80 (10.47)	15.35 (11.68)
Diamond	13.18 (8.48)	13.38 (9.06)

For each model, a deviance score was computed on participants' answers to the conditional probability questions. This was done by averaging the absolute value of the difference between the judged conditional probability and the expected conditional probability for the sixteen questions. These scores are shown in Table 2. A similar 4 (model) x 2 (condition) mixed Analysis of Variance was performed on these scores. A main effect of model was found: overall, deviance on the conditional probability questions differed among the four models: $F(3, 138) = 7.22$, $p < .001$. A main effect of condition was not found. No significant interactions were found. Simple effect analysis revealed that the deviance score on the chain with A->D model was greater than any of the other three models: $t(47) = 3.06$, 2.96, and 3.49, in contrast with the chain, common effect and chain, and diamond models respectively, all p-values $\leq .005$.

Did learners' error on the conditional probability questions reflect their ability to recognize causal structure? If learners are building accurate causal models, then responses to the conditional probability questions should be predictive of accuracy on the causal structure questions. To examine this, four hierarchical regressions were performed; accuracy on the causal structure questions for each model was the dependent variable. First, condition was factored into the model. Next, deviance scores on the other three models were factored in. This was done to control for participants' individual differences in the deviance scores. Finally, the deviance score for the model in question was factored in. This score contributed to the variance in the accuracy on the causal structure questions for the chain and the common effect and chain models: $\Delta r^2 = .110$ and .048 respectively, $F(1, 42) = 7.22$ and 3.99, both p-values $\leq .05$.

Discussion

Learners who were able to observe the results of their own interventions were better at recovering the causal structure among a set of events than learners who observed the results of interventions they were forced to make. This result parallels Sobel (2003), who found that observing the results of another's interventions resulted in worse causal learning than observing the results of one's own interventions. In both Sobel's (2003) observation of intervention condition and the present fixed intervention condition, learners were unable to test their own causal hypotheses. Even though learners might have observed critical information about conditional independence and dependence, learning in these circumstances was impaired compared with observing the results of one's own interventions. In fact, subsequent analysis revealed that all the participants in the intervention condition generated data that presented them with the specific conditional independence and dependence information necessary to learn each causal structure via constraint-based algorithms. Thus, all learners were given the relevant conditional independence and dependence information necessary to learn the causal structure. However, learning was facilitated when the learner controlled when they were given that information.

This experiment was motivated by two concerns with the procedure used by Sobel (2003). First, learners who observe the results of their own interventions might have simply been more involved with the task of learning than learners who observe another's interventions because of their ability to act on the system. Further, the benefit of observing the results of one's own interventions might by due to learners relying on kinesthetic learning skills. These two possibilities are inconsistent with the present data.

These data seem inconsistent with the hypothesis that constraint-based algorithms of causal structure learning best account for human learning. However, it is possible that constraint-based methods could be modified to account for these data. We suggest that Bayesian accounts of causal structure learning are at least more qualitatively consistent with these data than other approaches to causal structure learning. This is clearly a topic for future research.

One question that must be addressed is whether learners explicitly engage in hypothesis testing, or are relying on processes that are more implicit. Several researchers (e.g., Kuhn, 1989) have argued that children and in some cases, even adults lack the ability to design experiments that explicitly test causal hypotheses. Such deficits in scientific reasoning indicate that adults and children might lack the metacognitive skills necessary to design unconfounded experiments or interventions that discriminate among causal hypotheses.

Steyvers et al. (in press), however, demonstrated that adult learners would usually make a single intervention that offered them the most information to discern among various causal structures consistent with the data already observed. They proposed that there might be a difference between implicit causal knowledge, which is used in this kind of learning situation, and explicit causal knowledge, which was tested by researchers investigating scientific reasoning. For instance, participants in both the present experiment and in Steyvers et al.'s (in press) experiments were not asked to justify their responses, nor were they asked to reflect on the strategies they used to garner their knowledge. Learning through a Bayesian algorithm does not require explicit verbal representations of those hypotheses (see also Gopnik et al., in press).

To conclude, researchers in causal learning have been seeking a way to instantiate algorithms for learning causal structure. We believe that causal learning is better instantiated by the causal graphical model framework, particularly because of its ability to account for data from interventions (see Gopnik et al., in press). The present experiment represents an ongoing effort to investigate algorithms within that framework (see also Sobel, 2003). This line of research is more consistent with the qualitative predictions of Bayesian algorithms. Future work, however, must specify exactly what the nature of these algorithms is and how they might be represented in the brain.

Acknowledgments

We wish to thank Emily Blumenthal, Kathryn Noe, and Sara Yerry for their assistance in data collection, and Amy Hoff, David Lagnado, Sean Stromsten, and Josh Tenenbaum for insightful discussion and comments.

References

Furuhata, H. (1999). Traditional, natural and TPR approaches to ESL: A study of Japanese students. *Language, Culture, and Curriculum, 12,* 128-142.

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

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (in press). Theory formation and causal learning in children: Causal maps and Bayes nets. *Psychological Review*.

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94*, 211-228.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review, 96,* 674-689.

Lagnado, D. A., & Sloman, S. (2002). Learning causal structure. *Proceedings of the 24th annual meeting of the Cognitive Science Society.*

Murphy, K. P. (2001). *Active learning of causal Bayes net structure*. Technical Report: Department of Computer Science, University of California, Berkeley.

Pearl, J. (2000). *Causality.* New York: Oxford University Press.

Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General, 130*, 323-360.

Schulz, L. E. (2001). *"Do-calculus": Adults and preschoolers infer causal structure from patterns of outcomes following interventions.* Paper presented at the 2001 meeting of the Cognitive Development Society, Virginia Beach, VA.

Sobel, D. M. (2003). *Watch it, do it, or watch it done: The relation between observation, intervention, and observation of intervention in causal structure learning.* Manuscript submitted for publication, Brown University.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (in press). Inferring causal networks from observations and interventions. *Cognitive Science.*

Tenenbaum, J, & Griffiths, T. L. (2002). *Theory-based causal inference*. Proceedings of the 2002 Neural Information Processing Systems Conference.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20,* 273-281.