

Use of External Representations in Reasoning about Causality

David L. Mason

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
Under the Executive Committee
Of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2013

ABSTRACT

Use of External Representations in Reasoning about Causality

David L. Mason

This research investigated if diagrams aid in deductive reasoning with formal causal models. Four studies were conducted exploring participants' ability to discover causal paths, identify causes and effects, and create alternative explanations for variable relationships. In Study 1, abstract variables of the causal model were compared to contextually grounded variables and causal models presented as text or diagrams were compared. Participants given abstract diagrams did better in most tasks than participants in the other conditions, who all did similarly. Studies 2 and 3 compared causal models expressed in text to diagrammed causal models, and compared models using arrows to models using words when connecting variables. Participants who had arrowheads replaced with words made more errors than participants in other diagram conditions. Diagrammed causal models led to better performance than did other conditions, and there was no difference between different text models. Studies 4 and 5 tested the hypothesis that predictive reasoning (from cause to effect) is easier than diagnostic reasoning (from effect to cause). The two studies did not find any such effect.

Table of Contents

List of Tables	iii
List of Figures	v
Acknowledgments.....	vi
Dedication	vii
Chapter 1 – Causality and Diagrams	1
Diagrams and causality	2
Features to explore in causal diagrams	3
Chapter 2 – Literature Review	7
The Nature of Causality	7
Formal Computational Models of Causal Reasoning	10
Informal Causal Reasoning	11
The Role of Diagrams in Causal Reasoning	13
Future Avenues of Exploration	17
Study 1: Content	23
Methods	24
Results	29
Discussion	34
Study 2: Connections	38
Methods	38
Results	43
Discussion	47
Study 3: Replication of Connections Study	50
Methods	50
Results	55
Discussion	58
Study 4: Asymmetries in reasoning about causes and effects	60
Methods	61
Results	67
Discussion	72
Study 5: Replication of Asymmetry Study	76

Methods	76
Results	81
Discussion	86
General Discussion	88
References.....	96

List of Tables

Table 1.	Questions and answers for the Abstract conditions of Study 1.....	28
Table 2.	Median response time (in seconds) by condition for Study 1 Queries.....	30
Table 3.	Total score: descriptive statistics by condition.....	31
Table 4.	Mean proportion-correct score for each Query by condition.....	32
Table 5.	Study 2 tasks and answers.....	42
Table 6.	Median response time (in seconds) by condition for Study 2 Queries.....	43
Table 7.	Descriptive statistics for all four conditions.....	44
Table 8.	Mean proportion-correct score for Query by condition.....	45
Table 9.	Average of correct and incorrect answers by condition.....	47
Table 10.	Study 3 tasks and answers.....	54
Table 11.	Median response time (in seconds) by condition for Study 3 Queries	55
Table 12.	Descriptive statistics for all four conditions	56
Table 13.	Mean proportion-correct score for Query by condition.....	57
Table 14.	Study 4 tasks and answers.....	66
Table 15.	Median response time (in seconds) by condition for Study 4 Queries.....	67
Table 16.	Descriptive statistics for all four conditions	68
Table 17.	Mean proportion-correct score for Query by condition	69
Table 18.	Mean proportional correct scores for Cause and Effect Queries by condition...	71
Table 19.	Study 5 tasks and answers.....	80
Table 20.	Median response time (in seconds) by condition for Study 5 Queries	82
Table 21.	Descriptive statistics for all four conditions	83

Table 22.	Mean proportion-correct score for Query by condition	84
Table 23.	Mean proportional correct scores for Cause and Effect Queries by condition....	86

List of Figures

Figure 1.	The four conditions for displaying the causal model in Study 1.....	25
Figure 2.	The example models for Study 1.....	26
Figure 3.	Total Score by condition.....	31
Figure 4.	The results of the 2x2 ANOVA with the data from the Effect Query removed.....	34
Figure 5.	The four conditions for displaying the causal model in Study 2.....	39
Figure 6.	The example models for Study 2.....	40
Figure 7.	Total score by condition.....	45
Figure 8.	The four conditions for displaying the causal model in Study 3.....	51
Figure 9.	The example models for Study 3.....	52
Figure 10.	Total score by condition.....	56
Figure 11.	The four conditions for displaying the causal model in Study 4.....	63
Figure 12.	The example models for Study 4.....	64
Figure 13.	Total Score by condition.....	69
Figure 14.	The four conditions for displaying the causal model in Study 5.....	77
Figure 15.	The example models for Study 5.....	78
Figure 16.	Total Score by condition.....	83

Acknowledgments

As embodied by the models that so vexed my participants, bringing this dissertation to completion involved a vast network of collaborators, of which I was but a small piece.

Primarily I thank my advisor Jim Corter. Under his tutelage I have grown more than I ever could have imagined. He was everything I could have ever hoped for in an advisor: patient, dedicated, accessible, meticulous, and instructional. He was a mentor in the truest sense and inspires me to live up to the amount of faith he invested in me.

I also want to thank Barbara Tversky, whose expertise is only surpassed by her grace. She too was a great source of my growth. Her experience and insights were instrumental in my development as a scientist.

Aaron Pallas was great as the chair of my committee. His genial contributions were important in getting through the stresses of all my defenses. Steve Peverly and Sharon Schwartz were wonderful to serve as my committee examiners and provided fantastic feedback.

I also want to acknowledge all the people who contributed in other ways to effectuating this dissertation. Boyd Richards for his support, guidance, and recommendations. Herb Ginsburg and John Black for their respective roles in bringing me to TC. Everyone in the cog lab—Beetle, G-Unit, Sizeable Ted, Poppa Cap, Chunk, TNT, Caretaker, Schmitt, Half-Man Half-Amazing, Zartan, Twofer, and Tom Selleck—for all the talks, insights, revelry, food, and camaraderie.

Lastly, thank you to my family. Both sides have been nothing but supportive and necessary stalwarts during difficult times. A special “thank you” to my parents for one heckuva influential plane ticket. To my wonderful wife and son, thank you. You were both amazing. We were all in this together and you share equally in the accomplishment.

Dedication

*This work and everything it represents is dedicated to Heather, the foremost example of all that
is good in the world;
and to Rhys, who embodies that legacy*

Chapter 1 – Causality and Diagrams

Causal inferences affect future behavior. This is because a causal model can be used to explain past events and predict future outcomes. People assess the strength of the relationship between two events, interpret that relationship using implicit or explicit causal models, and then use that information to predict future events. Without this ability to reason causally, relationships between events would appear arbitrary, and planning for a future event would be impossible. For this reason, the nature of causality and how people reason about it has been addressed for millennia and in diverse fields such as philosophy (Bacon, Campbell, & Reinhardt, 1993), psychology (Piaget, 1930), physics (Bohm, 1957), economics (Zellner, 1988), medicine (Stehbens, 1992), statistics (Freedman, 2007), and computer science (Dean & Kanazawa, 1989).

These attempts in different fields and at different historical time periods to understand causality and causal reasoning have had similar aims and emphases. All have sought to address how and why events are related and how people perceive those relationships. Often this is done through the use of formal mathematical methods.

It is useful to consider causal reasoning as involving two subcategories of reasoning. There is the reasoning that occurs as a person makes efforts to assign causal properties to events. An example would be feeling heat when passing a hand over a flame and determining that the fire caused the heat. This inductive process serves the purpose of creating a causal model to use for predictions in the future. These deductive predictions from a causal model comprise the second category of causal reasoning. Having already determined that fire causes heat, if a person wanted to be warm, they could use their causal knowledge to start a fire that would satisfy their need. Causal reasoning often involves a series of interleaving instances of inductive and deductive reasoning as one creates a causal model, puts it to the test, and then

makes the necessary adjustments to hone the model. For example, a tutor may reward a pupil with a quick video game break during a lesson. When the student comes back refreshed, the tutor may reason that video games cause better performance. But after observing the same results after a snack break, the tutor may revisit the original causal model and modify it so that instead of crediting video games with the student's performance, now the more general concept of taking breaks is thought to cause improved performance.

The question of how a person actually makes such inductive and deductive causal inferences is still a matter of some debate. The modern conception of causality is often attributed to David Hume (1739/2000) who proposed an associative theory of causal induction. This idea that people infer causal attributions based on the covariance of two or more events was the dominant theory for the next 250 years (Mill, 1843/1974). More recently, theories have begun to recognize other sources people use for causal inference, such as prior knowledge or experimentation (Ahn, Kalish, Medin, & Gelman, 1995; Cheng, 1997; Pearl, 2000).

Diagrams and causality

As the theories explaining causal reasoning have become more sophisticated, so have the methods used to represent them. Causal diagrams (e.g. directed acyclic graphs, Bayesian networks) are frequently used to express a causal model. Diagrammatically representing complex information has a large literature justifying its use (Larkin & Simon, 1987; Tversky B. , 2005) but always with the caveat that subtle design choices can lead to systematic errors and biases (Tufte, 1983). And although diagrams are ubiquitously associated with causal models, little research has been done on the actual cognitive and performance advantages of using causal diagrams, if indeed, any exist.

Research on best practices for causal model representation might focus on two scenarios: how people employ external representations to construct, modify and use causal models themselves, and how they might use external representations to convey a causal model to others. One instantiation of the latter scenario involves situations where experts are doing the inductive attribution work and then communicating complex findings to an untrained audience. Examples include a climatologist presenting findings regarding climate change to a legislator, an organizational psychologist explaining results to a company's board of directors, a sociologist presenting educational research to a superintendent, or a journalist writing to her readers about healthcare. What the recipients of these models do with the findings, through deductive reasoning, has the potential for shaping public policy, business practices, or general lifestyle. It is therefore important to prevent any break in communication between what the causal model is meant to express and what is actually perceived.

Features to explore in causal diagrams

It may be helpful to consider how use of a causal model might be affected by three aspects of the model: the content of the model, diagrammatic elements of the external representation, and cognitive biases that affect use of the model.

Content

Research has shown that people reason about abstract material differently than they do about socially grounded material (Cosmides, 1989; James, 1975; Schwanenflugel & Shoben, 1983) often demonstrating more success with the grounded material. There is not a consistent advantage for contextual material over abstract material, however. Familiarity may also lead to biases that can inhibit correct interpretation of the specified model. This may occur if the model identifies relationships that seem implausible based on the reader's pre-established mental model

(Easterday, Aleven, Scheines, & Carver, 2009; Stroop, 1992) or if the model introduces irrelevant information that can distract from the task (Day, Manlove, & Goldstone, 2011; Kaminsky, Sloutsky, & Heckler, 2008).

Because formal causal models are created to represent real-world findings, context is an integral component of the model. Provided the model represents relationships that can be deemed plausible, it would seem that a model posed in a specific context would be processed more easily than an abstract model. However, the primary purpose of the causal model is to represent the existence and directionality of causal relationships between variables, therefore the additional information provided by grounded variables might serve only to obscure those relationships, thus detracting from model-based inference tasks.

Diagrammatic elements

A causal model is relatively simple in its construction. It consists of variables and the connections between those variables. A diagram adds the components of symbolic connections between variables (arrows) and using space to arrange the variables in a way that a causal model presented as text cannot. Previous research has shown that participants who were presented with diagrammed causal models were able to answer questions about the model more successfully than participants who received the model in text form (Corter, Mason, Tversky, & Nickerson, 2011). Even though the two models are informationally equivalent, they appear not to be computationally equivalent. That is to say, a diagrammed visualization seems to allow for easier interpretation compared to a text-based representation. This advantage in performance for diagrammed models likely stems from some combination of the advantage of an explicit spatial array to represent causal priority and the use of arrows that both connect variables and indicate directionality of specific causal relationships.

Arrows are rich with symbolic meaning (Heiser & Tversky, 2006) and may be a significant aid in interpreting causal models. In a diagrammed causal model, arrows represent the asymmetrical temporal relationship of a cause and an effect. However, words share similar properties to arrows in that a word is read from left to right and often has asymmetrical properties as well. For example, the word “affects” always means that the noun in the subject part of the sentence is the agent acting upon the noun in the predicate part of the sentence. There is no ambiguity that the predicate is actually acting on the subject.

It would appear that given the similarity between words and arrows, any advantage in performance for participants with diagrams over participants with text could be reasonably attributed to the spatial organization of diagrams. However, even though the connections between variables (arrows and words) are similar in purpose, there is not yet any evidence that they result in similar performance. The elimination of connection type as a contributing explanation for performance differences would bolster the case for spatial array as the primary explanation for why diagrammed causal models improve reasoning.

Cognitive biases and errors

The final category to explore is the inherent biases that people bring when reasoning about causality (Kelley, 1973; Maldonado, Jimenez, Herrera, Perales, & Catena, 2006). One of these potential biases is the idea that there is a difference between reasoning about the effects of causes and the causes of effects (Mill, 1843/1974). There are several reasons to expect an asymmetry between the ease or accuracy of reasoning from cause to effect (predictive inference) and from effect to cause (diagnostic inference). These reasons include causal features being more salient (Ahn, 1998; Ahn, Kim, Lassaline, & Dennis, 2000; Tversky & Kahneman, 1980),

psychological essentialism (Medin & Ortony, 1989), and temporal consistency (Hume, 1739/2000; Tversky & Kahneman, 1980).

There is reason to doubt a universal bias towards predictive inference however. Research has shown that several factors, e.g. the amount of available evidence for causes and effects, are more instrumental in determining which inference is easier (Fernbach, Darlow, & Sloman, 2011). Research in this area has been conducted by presenting participants with either simple, dichotomous text or verbal problems. Framing the question as a complex model or by using a diagram is as of yet, unexplored.

To summarize, much research in cognitive science and related fields has examined the nature of inductive causal reasoning. In formal methods for causal modeling, diagrams have been used almost since the inception of the techniques (Wright S. , 1920). However, little research has been done on whether diagrams are actually useful for causal representation and causal reasoning. Nor has there been much research on how non-experts deductively reason from explicit causal models. Research addressing the type of content of a model, diagrammatic elements of the model, and pre-existing biases that affect people's interpretations of causal models, should be useful to achieve better understanding of the utility of using diagrams to represent causal models.

Chapter 2 – Literature Review

The concept of causality has been studied in many disciplines. The nature and definition of causality is of great interest to philosophers examining the ontological nature of the relationships between events. Scientists in many sub-disciplines wish to determine the extent to which an event causes an effect in order to make predictions with greater accuracy.

Programmers need to understand how people learn causal relationships so they can replicate that learning in machines. Many cognitive scientists are interested in causal reasoning, to understand the thinking that people do in order to simply navigate their lives.

It is useful to consider causal reasoning as consisting of two categories, or types, of reasoning. The first type, inductive reasoning, is where people take evidence and create (either explicitly or implicitly) a causal model. The second type, deductive reasoning, is the process of inference from an already-established causal model. These terms will be referred back to frequently throughout this review.

The Nature of Causality

The notion of causality has long been of interest to philosophers and scientists alike. In Western philosophy, the first recorded analytic treatment of causality can be traced back to Aristotle (Aristotle, n.d./2004). His view of causality is more closely aligned with our current idea of ontology in that he was searching to explain an object's existence (its cause for being) rather than identify how one event relates to another. He described the cause of every object as being composed of four causal elements: the material cause, formal cause, efficient cause, and final cause. The material cause of an object is that of which it is made. For example, the material cause of a chair is the substance (wood, upholstery, etc.) from which it is constructed. The formal cause is the form of the object, i.e. a chair is a chair because it is shaped like a chair.

The efficient cause is the process of change that resulted in the object. That is to say because the materials went through a chair-making process; the chair's efficient cause is that process. Lastly the final cause is the ostensible purpose of the object. Because one sits on the object, its purpose is to be a chair. The combination of these four causes is what gives an object its explanation of being.

The Stoic philosophers are to be credited with our modern conceptualization of causality as a description of a deterministic relationship (that an effect would not have occurred without its cause). They dismissed most of Aristotle's other causal elements, arguing instead that all causation was efficient causation (Frede, 1987), that every event was an effect which necessitated having a cause.

Although the Stoics are responsible for changing the idea of a cause from a description of being to a description of a relationship, much of how the study of causality is approached currently can be largely attributed to David Hume (1739/2000). His approach was to focus less on the explanation of *why* something occurred but simply to focus on the explanation of *how* it occurred. He proposed that we are unable to actually perceive cause and so we inductively draw probabilistic conclusions based on observed covariation. What we call "cause" is merely a way of describing the temporal order of covarying events. It was Hume's rules about temporal contiguity and order that evolved the study of causality to the one that persists today.

Immanuel Kant (1781/1999) challenged Hume's assertions about the unknowable nature of causality by proposing an alternative explanation. He argued that Hume was too skeptical about causality's existence because he focused too much on empiricism as being objective truth. Contrarily, Kant believed that causality could exist because of his belief in objective truth and

that at some level, humans have an *a priori* concept of how objects and events relate to each other which we then use to structure our experience.

John Stuart Mill (1843/1974) advanced the idea that a singular cause was rare and that effects were the result of several partial causes. He also discussed negative cause, that the absence of something can itself be a cause. Mill developed several rules of causal logic, known informally as Mill's Methods, to explain appropriate causal induction. These rules were predicated on the notion of a cause being necessary and sufficient. That is to say, he proposed that a cause (i.e. the combination of necessary partial causes) would be sufficient to invariably produce a particular effect, regardless of other influences. Mill's work was instrumental in establishing the philosophy of science.

J. L. Mackie (1965) contributed to the philosophy of causality by proposing that an event can have multiple causes (contrast this with Mill's idea that one cause can be made up of several components). This distinction is important because the work of the aforementioned philosophers had supposed that causes were both necessary and sufficient to bring about an effect. Mackie explored this idea using the example of investigating a fire and concluding that an electrical short was the cause of the fire. There are several other crucial components to this scenario such as not having a proper sprinkler system and flammable material being near the spark, both of which were necessary for the fire to have occurred. Thus, in many situations, when people refer to the cause of something, they are actually referring to an event that by itself is *insufficient* to bring about the effect but nonetheless *necessary*, but only within the context of a larger condition (the causal field) that by itself is *unnecessary* but *sufficient*. This is known as an INUS variable, drawing from the acronym created by the previously italicized words. So because an electric

short is not the only way to start a fire, the electric short is both not necessary and insufficient unless there is a condition where all that is missing to start the fire is the electrical short.

Formal Computational Models of Causal Reasoning

Modern conceptions of causality follow in the tradition of Hume and Mill with the idea that causes are made up of material elements which can be studied. Under this assumption, many attempts have been made to move beyond merely logical representations of causality and establish a mathematical computational model that can be used to parse the degree of causal influence an event has on an effect, and to automatically generate deductive inferences.

The first attempts at such a theory were by Yule (1899) who tried to use the regression techniques developed by Legendre (1805) and Gauss (1809) to establish a causal link from the correlational data regarding the policy decisions of local representatives and the number of people on welfare. Ultimately Yule decided he was not able to make any causal assertions.

The next step in the evolution of causal representation can be traced to Sewell Wright (1918; 1921) a scientist working for the United States Department of Agriculture. His contribution grew from observations about data regarding rabbit growth that correlations might be misleading because they used overlapping variance in explanations. He developed the technique known as path analysis, which works by creating an *a priori* causal model of the given variables and then using regression techniques to estimate path coefficients to the proposed connecting causal links, thereby allowing a researcher to both measure relationships and (indirectly) to posit new ones.

Structural Equation Modeling is a general framework for fitting causal models, differing from path analysis by using maximum-likelihood estimation methods and by allowing for modeling of latent (unobserved) variables. The techniques of SEM were developed through

several fields simultaneously during the 1960s and 1970s, thus SEM does not have a single inventor although Sewall Wright is often considered to be its progenitor (Bollen & Pearl, 2012).

It is important to emphasize that any causal claims that stem from these methods are assumptions provided by the researcher. Path analysis techniques do not infer causality so much as assess the evidence for a researcher's claim of causality. Because the causal model is essentially in place in the mind of the researcher, these formal methods of representation are used to deductively reason about the variables and do not provide insight into how people inductively identify causal relationships.

Additionally, people do not reason mentally about causality using formal mathematical models. Rather most inductive and deductive causal reasoning, e.g. identifying what foods make one sick or finding appropriate retorts to a spouse, is done in an *ad hoc*, informal manner. The mechanisms for accomplishing such tasks are still unclear although several theories have been posited.

Informal Causal Reasoning

Hume's theory of strict associationism was a great influence on early behavioristic learning models in psychology and related fields. One of the most widely-cited of these models is the Rescorla-Wagner (R-W) model (Rescorla & Wagner, 1972). This classical conditioning model proposed that learning occurred when expectations were most disparate from outcomes. As someone becomes accustomed to observing a co-occurrence relationship between two events, the surprise diminishes and the relationship is considered to be learned. By adding the element of expectation, this model can account for unresolved issues (e.g. overshadowing and blocking, where a second stimulus is erroneously perceived to be weaker) that occur with a strictly associative view of learning (Miller, Barnet, & Grahame, 1995).

There remain however, several phenomena that the R-W model cannot account for, such as backward blocking (a retroactive reduction in causal strength of a second variable) and how the most salient stimulus is not necessarily considered to be causal (Buehner & Cheng, 2005; Cheng, Novick, Liljeholm, & Ford, 2007). These problems stem from the basic assumptions inherent in associationism. One way to resolve these problems is to incorporate the use of an *a priori* causal mechanism (Kant, 1781/1999), known now as a power or rule-based theory. In other words, the reason people do not say that a rooster causes the sun to rise or that Thursday causes Friday is because people have to have some reason to believe that the cause generated the effect rather than merely coexisted with the effect (Ahn et al., 1995). Patricia Cheng proposed a formal model that combines the associative and the *a priori* concepts without the flaws that plague either theory and called it the power probabilistic contrast model (Cheng, 1997; Cheng & Novick, 1990).

Another recent theory of formal causal inference is Causal Bayesian Networks (Pearl, 2000). This is a method that mathematically estimates the effect of confounding variables (by computing and reconciling appropriate conditional probabilities), thereby creating an *ad hoc* experimental condition – a simulated intervention as it were – that estimates the degree of causality. Such Bayesian Networks (BNs) have been argued to combine ideas from associationism (i.e., the conditional probabilities), rule-based theories (the set of included variables), and Mackie’s concept of INUS variables, and returns mathematically justifiable and useful results. One of the benefits of Bayesian analysis is that it bridges the gap between inductive and deductive causal reasoning (Lagnado, 2011).

Although causality at its most basic can be described simply as “A causes B” – and science attempts to minimize experimental variables in order to achieve that simplicity – rarely

do natural phenomena present such a clear manipulation. Previous efforts from cognitive science, statistics, and philosophy have documented a number of errors that people make in causal inference. These errors can be made both inductively (formulating an incorrect causal model on the basis of evidence) or deductively (using a causal model to draw incorrect inferences or predictions about events). An inductive error is one that occurs when trying to create a causal model based on evidence, such as an observed (spurious) correlation. For example, observing a correlation between owning a laptop and doing well on a mathematics assessment, and forming a theory that the former causes the latter (not considering other related variables such as socio-economic status) is an example of using spurious relationships among variables to induce an incorrect model. A deductive error is one that occurs when the model is in place but erroneously interpreted. One possible example of a deductive error would be an incomplete search for causal factors (a phenomenon known in the computational learning literature as “blocking”). A specific example is if a school administrator—presented with a causal model that showed the effect of SES on both use of laptops and math scores—were to ignore the effects of SES anyway, inferring the spurious correlation because laptops seemed to explain the math scores.

The Role of Diagrams in Causal Reasoning

Although causal models have become synonymous with diagrammatic representation (e.g. directed acyclic graphs, Bayesian networks), it is interesting to note that initial efforts in modeling causality did not use diagrams. In the paper on poverty and politicians, Yule (1899) drew only a table. In Wright’s first paper where he described path analysis (1918), he simply presented a table in which he computed partial correlations. A follow-up paper (1920) is where the familiar diagrammatic path analysis first appears. In Pearl’s (1985) first paper on the subject,

he did not represent Bayesian networks as a diagram. This observation highlights the fact that diagrams are not critical to these techniques.

However, to the modern user path analysis and Bayesian networks would seem incomplete if an accompanying diagram did not appear. So what is it about diagrams that make them ubiquitous in studies on causality? Is their appeal purely esthetic? Do they simply create an interface that simplifies an otherwise overly complex idea? Can diagrams contribute to reasoning in a way not otherwise easily accomplished? Or are they merely useful as external memory which can free up cognitive processes for reasoning?

Pearl (2000) stresses the importance of using diagrams to reason about the nature of the model. Diagrams are not bound by the spatial constraints of sentential text. Diagram designers can take advantage of this fact, grouping relevant data together in a way as to minimize searching (Larkin & Simon, 1987). Sometimes perceptual cues are present that can increase the amount of information being presented. For example if a line bisects another line at a 30-degree angle, it is perceptually salient that the neighboring angle is (roughly) 150 degrees. The visualization also facilitates a host of other inferences not readily apparent from the equivalent text description (Larkin & Simon, 1987; Scaife & Rogers, 1996). These organization advantages mean that diagrams can aid mental computation because a diagram can present equivalent information to text but in a format better suited for inference (Larkin & Simon, 1987). Diagrams can help make implicit possibilities explicit. In one study (Bauer & Johnson-Laird, 1993) participants were given a reasoning problem where the diagram version was arranged to look like puzzle pieces. This cue helped participants figure out the problem both faster and more accurately than text. The text articulated the same limitations demonstrated by the puzzle pieces, but that did not translate to equivalent success by the problem solvers. Diagrams can also

accentuate relevant features, such as in caricatures or in maps (Mauro & Kubovy, 1992; Tversky et al., 2007). Creating diagrams is a “deep processing” task that can improve understanding: In one experiment, students who created their own diagrams performed better on subsequent tasks – one being understanding causal relationships – than students who summarized the same content through text (Gobert & Clement, 1999).

Empirical Studies on Diagrams and Causality. Despite a great interest in improving causal reasoning and a large body of literature on the usefulness of diagrams in general, only a few studies have been done specifically on use of causal diagrams and their benefits for causal reasoning. One study found no difference in performance between participants who studied three pages of text explaining a causal model and participants who were presented with a diagrammatic model (McCrudden M. T., Schraw, Lehman, & Poliquin, 2007), when both groups were tested on their memory of the relationship. The absence of a difference in outcomes was viewed as a positive because of the increased efficiency of the diagrammatic condition. However it was not clear whether causal diagrams were beneficial at the encoding or retrieval stage (or both). In a subsequent experiment the authors studied whether explicit lists worked the same as causal diagrams (McCrudden, Schraw, & Lehman, 2009) and found that when providing an extra study tool – either in the form of rereading the text, getting a list of causal steps, or a diagram of causal steps – the list and the diagram conditions were generally equivalent and both were better than rereading the text. They hypothesized that the findings might differ for experiments that were to employ causal models more complex than the simple linear one they used (i.e., “A causes B which causes C” and so on).

Other studies (Easterday, Aleven, & Scheines, 2007; Easterday, Aleven, Scheines, & Carver, 2009) have found that participants who were given a diagram *and* textual explanation

did better in reasoning about a public policy problem than those who had only the text or those who had the text and then constructed their own diagram. However, on a subsequent related task where they were again only presented text, those who had previously *constructed* diagrams did better than the other conditions even though the second task did not involve diagram construction. Those who had previously only seen diagrams and text did better than those who had only seen text.

However, in those studies, the conditions were not entirely equivalent, because people in the text condition received less information than in the other conditions. This confound occurs in other studies as well (see McCrudden et al., 2007; Langley & Morecroft, 2004), making conclusions difficult to generalize.

Some previous research has attempted to assess the effects of using diagram with a more complex causal model, keeping informational content equivalent between conditions (Corter et al., 2011). In one experiment, participants were given a causal model consisting of five nodes that were interconnected in nine ways. Compared to those who saw a sentential (text) representation of the same model, participants were both more accurate and faster at completing the task. Another experiment used an even more complex model (six nodes, ten connections) and tested whether the depicted direction of causal flow in the diagram mattered. Participants with diagrams were again faster and more accurate. But this research also showed that a diagram with causal flow depicted as from left-to-right was processed faster than a diagram with causal flow reading from right-to-left. The right-to-left diagram had similar performance outcomes but the time for task completion was almost twice as long as the left-to-right diagram condition.

Based on these findings, it appears that diagrams aid reasoning about causality better than text although the reason for that is unclear. Additionally, it appears that aspects of the diagram

itself can be designed to further facilitate reasoning. The exact components of a diagram that are relevant and useful have yet to be categorized.

Future Avenues of Exploration

Because of the emerging nature of studying how diagrams affect causal reasoning, there are several areas within which to explore these questions. Lines of research that may prove fruitful are to examine the role several factors, specifically the content of the causal model, the structure of the causal model, and cognitive biases that can affect how accurately inference proceeds.

Content

Diagrams used with causal models in real applications are by their very nature grounded in a real-world context. Research has shown that this grounding can be beneficial (Cosmides, 1989; James, 1975; Schwanenflugel & Shoben, 1983), with people showing better performance outcomes when reasoning with concrete material. One possible explanation for this finding is described by the Dual Representation Model (Schwanenflugel & Shoben, 1983), which assumes that specific domain content / context encourages both visual and verbal representations, whereas abstractions only activate verbal representations. Another possibility is described by the Context Availability Theory (Kieras, 1978), which assumes that familiarity with the context provides additional information which is unavailable for abstract concepts, and which may aid inference. Therefore it would be informative to include problems in our research that involve socially- and pragmatically- grounded relationships, and to examine if reasoning in these types of problems differs from those using abstract relationships. A participant's previous familiarity with the real-world domain tapped by the variables in the contextually grounded model may reduce the

cognitive effort necessary for keeping track of those variables and their relationships (Ericcson, Chase, & Faloon, 1980), and may reinforce causal inferences via experience-based learning.

However, research has also shown that in some cases context can actually inhibit performance. This seems to happen when the task is poorly understood (Cummins, Kintsch, Reusser, & Weimer, 1988; Geary, 1994), when the context is counterintuitive (Easterday, Aleven, & Scheines, 2007; Stroop, 1992), or when the context introduces superfluous information that distracts from the problem (Day, Manlove, & Goldstone, 2011). As described in the previous section on the nature of causality, causal inference is often treated as a contextually grounded logic problem. Thus, context may complicate deductive reasoning. This may be one of the reasons that the techniques of causal modeling are usually taught using abstract variables.

Structure

Causal models are interesting in that they convey both structure and function. The structure of the model is defined by its makeup of variables and their relations. The functional aspect of the model is that those relations indicate effects or “operations”, i.e. that one variable causes another. As such, causal models seem to be ideally suited for diagrammatic representation. However, research has shown that diagrams can create misinterpretations if they not designed correctly (Tufte, 1983).

In comparing causal diagrams to causal text, there are few structural differences to explore. They both are made up of variables and both use links to connect those variables. The links are similar in purpose yet different in representation. Arrows, typically used in causal model diagrams, have a rich symbolic history, invoking several hundred possible meanings

(Horn, 1998). Their use may be an important factor in what makes diagrams easier to interpret than text.

The diagrammed model also introduces a component not present in the text-based sentential model, the use of space. Larkin and Simon (1987) make the point that the computational cost of searching for information is less when searching in a diagram than when searching text containing equivalent information because of the way information can be organized. Additionally, the cognitive effort for inference can be lessened as well. In a causal model, the information about variables and their interconnectedness is immediately perceptually apparent due to its spatial organization, but not when presented in sentential format.

The cognitive effects of these specific components of diagrams, arrows and the spatial layout of causal models are issues in need of further exploration to determine the extent of their influence on causal inference and whether manipulation of those elements might lead to errors.

Cognitive bias

People develop the notion of causality at an early age (Cohen, Amsel, Redford, & Casasola, 1998; Goswami & Brown, 1990; Leslie & Keeble, 1987). Because even simple forms of causal reasoning can be adaptive, there is adaptive value in developing cognitive heuristics that aid (but may sometimes hinder) causal reasoning. People seem to assign causality spontaneously to a relationship between co-occurring events (Wong & Weiner, 1981), which suggests that the causal mechanism may sometimes go awry (e.g., in the formation of superstitions). There are many causal errors that might be investigated. We focus on possible biases in reasoning deductively from cause to effect versus from effect to cause.

The idea that reasoning from cause to effect (predictively) and reasoning from effect to cause (diagnostically) may not be equally difficult has been around at least since the 1800s (Mill,

1843/1974). Recent research indicates that predictive reasoning is often easier (Ahn, 1998; Ahn et al., 2000; Tversky & Kahneman, 1980). One possible explanation of why causes and effects have asymmetrical status involves the temporal position of causes and effects. As Hume (1739/2000) observed, a cause must precede its effect. Our experience in the world is that time proceeds linearly, with effects never occurring before causes. When reasoning about causes and effects, it may be more intuitive to think in chronological order.

Diagnostic reasoning also has a fundamentally different structure than predictive reasoning (Pearl, 1988; Waldmann & Holyoak, 1992). In predicating and confirming a cause and effect relationship, both predictive and diagnostic reasoning must be aware of what effects a cause must have. But in diagnostic reasoning, one must also be aware of competing causes. For example, knowing that an acquaintance was driving drunk, it would be easy to predict a subsequent car accident. However, knowing that an acquaintance was in a car accident offers little evidence of previous inebriation as the accident could have been caused by a myriad of other factors.

Another factor to consider is that in a cause-effect relationship, the cause is the *agent* of change. The effect is the passive result. This is reflected in the typical order of sentence construction, where the subject acts upon the predicate. Reversing the order and placing the acted-upon entity as the subject is considered a special (less effective) case, the passive voice. In general, the actor is awarded a preferred position of primacy in language (Ferreira, 1994). These related ideas may bear on interpreting causal models. A cause both occurs before and brings about its effect. Thus the mental representation of a cause may reflect both its temporal position and its status as agent. The process of interpreting a causal model, especially a complex one, may be aided by framing questions in such a way as to assure the cause-effect relationship is

aligned with its most intuitive mental representation. That is to say, it may be easier to reason about X causing Y than to reason about Y being caused by X.

Another possible explanation for the cause-effect asymmetry is the idea of *psychological essentialism* (Medin & Ortony, 1989). If I use a hammer as a paperweight, is it not a hammer anymore? The debate revolves around the concept that the hammer contains an essence allowing it to retain its “hammer-ness” even when not being used in its traditional sense. This essence—whether form, purpose, or intention—is used to produce or limit inferences about the object. Cause may serve the same “essence” function when we categorize an event, at least more so than an effect does (Ahn et al., 2000). That is to say, causal features may be more salient (essential) because they are used to infer and predict.

Summary

Much research has been conducted on the nature of how people reason causally. Furthermore, there are visual methods typically used for representing causal relationships, primarily directed acyclic graphs or DAGs. However, little research has been done on how well people reason using those visual representations and methods. Causal model diagrams have three categories that should be addressed in order to determine the effect of visualizing the model on causal inference. The first issue to address is the actual content of the model. Causal models are usually posed in terms of real-world variables. But the effect of specific domain context on inference depends on various aspects of the problem in question. It remains to be seen whether causal reasoning is generally helped or hindered by specific domain context. The second issue is that diagrams are generally considered to be useful in easing cognitive burdens for processing complex information. But little empirical evidence exists to support this belief, or that could be used to design more effective visualizations. Careful exploration of the features of a causal

diagram can shed light on the efficacy of different visual features of a causal model diagram.

The third issue is that of cognitive biases. People typically engage with a causal model using their already established heuristics for reasoning about causality. These heuristics may be innate, but more likely arise through extensive informal experience in normal life. Even with a formal causal model intended for reasoning about variables, problem characteristics may trigger use of a heuristic that could lead to systematic biases that can interfere with correct reasoning. It is the goal of the following studies to explore these issues not previously addressed in the literature.

Study 1: Content

Study 1 had two goals, the first of which was to compare how well people reason using causal models presented as text compared to models presented as diagrams. Previous work (Corter et al., 2011) had shown diagrams to have a positive effect on successful interpretation of a causal model.

The second goal was to compare how the content of a causal model affects successful causal inference. Specifically, Study 1 examined whether content comprised of abstract variables leads to more accurate inferences compared to content expressed in terms of variables grounded in a real-world context. Generally, contextually grounding variables benefits reasoning when the additional information provided by context contributes to the task at hand (Baranes, Perry, & Stigler, 1989). One reason for this effect is that the burden on working memory is decreased (Ericcson, Chase, & Faloon, 1980), which allows more cognitive resources to be devoted to reasoning rather than memory. Additionally, people are thought to develop pragmatic reasoning schemas through personal experience which may aid correct reasoning, e.g. the “permission” schema (Cheng & Holyoak, 1985; Lehman & Nisbett, 1990).

Conversely, familiarity with a given context may also lead to biases that can inhibit correct interpretations. This may occur if the context of the relationships seems implausible based on the reader’s pre-established mental model (Easterday, Aleven, Scheines, & Carver, 2009; Stroop, 1992) or if the model introduces irrelevant information that can distract from the task (Day, Manlove, & Goldstone, 2011; Kaminsky, Sloutsky, & Heckler, 2008).

Formal methods for estimating and computing with causal models (e.g., path analysis or Bayesian networks) are often taught using abstract contexts, but typically used to represent real-world (i.e. contextually grounded) variables. Study 1 presented participants with models

containing either abstract or concrete variables for the purpose of ascertaining whether the additional information provided by context was an aid or a hindrance to reasoning in a purely deductive inference task.

Methods

Participants

A total of 240 participants were recruited from Amazon's Mechanical Turk (MT). MT is a website that manages a marketplace for employers and workers in Human Intelligence Tasks (HITs), typically tasks which present difficulties for artificial intelligence solution. These tasks range from proofreading translations and tagging images to complex psychological experiments and may require anywhere from a couple of seconds to several minutes to complete. Monetary compensation varies for these HITs depending on the time investment.

Because the present study involved interpreting English text, 67 non-native English speakers were excluded from the analysis. Thirteen more participants were excluded due to not completing the task. This left 160 remaining participants across the four conditions.

The participants were 57% male. Most participants reported having had some experience in higher education with over 90% reporting having attended some college and 37% reporting having attended some graduate or professional schooling. About 42% of participants who reported college majors were from mathematical/engineering/computer science programs. About 45% of participants indicated having taken only one or two classes on statistics and 50% reported never having taken a statistics class. The average age of participants was 31 years with a range from 18 years to 67 years.

Stimuli

Participants were presented with one of the four causal models depicted in Figure 1. The models were identical in structure but varied in how the structure was expressed (either as text or a diagram) and whether the variables were abstract or concrete.

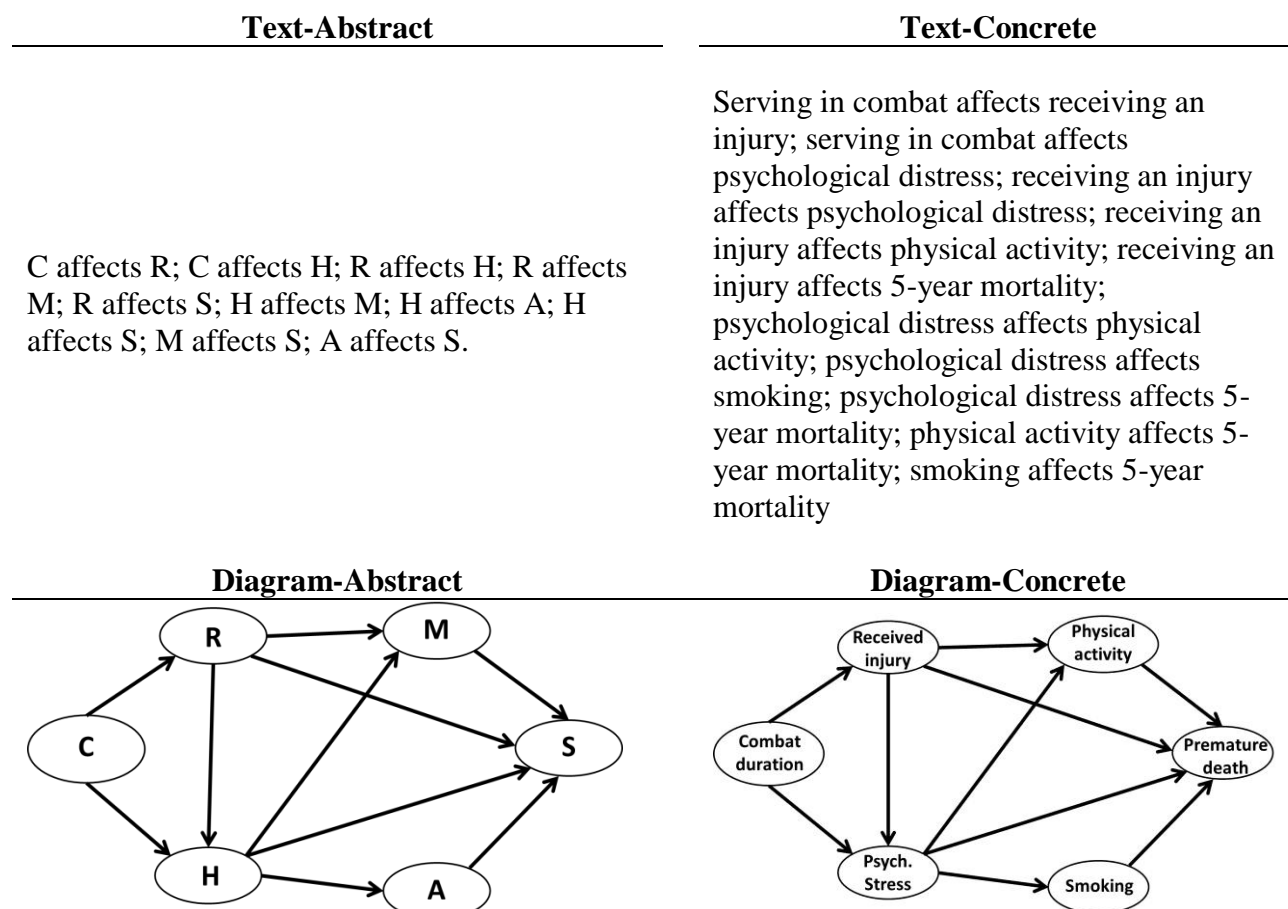


Figure 1. The four conditions for displaying the causal model in Study 1.

Procedure

This experiment was hosted online in Amazon's Mechanical Turk (MT). The experiment-administration code was written in HTML and JavaScript. Participants were paid \$1.00 to complete the task. Once participants accepted the task, they were randomly assigned to one of the four conditions and presented with a short introduction. They were allowed a maximum of

one hour from the time they accepted the task to fully complete and submit it. Following is the introduction for all conditions:

In causal modeling of a social science problem, researchers try to specify all the ways in which variables might influence each other.

For example, a researcher might assume that variables X, Y, and Z have the following causal relationships: X affects Y, X affects Z, and Y affects Z.

*In that case, X has a causal influence on Z in two ways. First there is a **direct effect** of X on Z (by assumption). Also there is an **indirect effect** of X on Z, because X is assumed to affect Y and Y is assumed to affect Z.*

To summarize, if you were given the following set of causal assumptions:

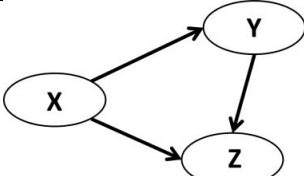
Text example	Diagram example
<i>X affects Y, X affects Z, and Y affects Z.</i>	 <pre> graph LR X((X)) --> Y((Y)) X((X)) --> Z((Z)) Y((Y)) --> Z((Z)) </pre>

Figure 2. The example models for Study 1

At this point, participants were presented with the causal model example in either text or diagram form depending on the condition to which they were assigned (see Figure 2). The instructions continued:

and you were then asked to write the direct and indirect effects of variable X on variable Z, you would write:

X affects Z.

X affects Y which affects Z.

After the introductory example, participants clicked the “proceed” button and saw the following task description and the target causal model in the Abstract conditions:

Now, assume that a researcher makes the following causal assumptions about a particular social science domain where variables C, R, H, M, A, and S are measurable aspects of people.

The researcher makes the following causal assumptions:

In the Concrete conditions, the introduction to the task replaced the list of abstract variables with descriptions of the concrete variables.

Now, assume that a researcher makes the following causal assumptions about the effects of combat on veterans who have returned home using these variables:

Combat Duration = length of time in a military combat zone

Received Injury = sustained an injury during combat that required hospitalization

Psych. Stress = score on a particular psychological assessment for veterans

Physical Activity = amount of exercise during 2 years after tour of combat

Smoking = amount of smoking during 2 years after tour of combat

Premature Death = Died earlier than average life expectancy

With the preceding text and one of the four models (see Figure 1) visible on the screen, participants were presented with the first of four questions. After they answered it, they pressed a button and the first question (along with their answer) disappeared and the second question appeared. Meanwhile, the model remained on the screen. This was repeated until the final question which, upon completion, directed participants to a new page asking for demographic information. From this page they submitted their entire task. Participants were not permitted to go back to a previous screen to view or change answers.

Outcome measures

Participants were asked four questions regarding various aspects of the causal models. With the exception of the variable names, the questions and answers were identical across conditions. The first question (Path Query) asked participants to identify all direct and indirect effects between two variables. The second and third questions (Cause Query and Effect Query respectively) asked participants to identify causes and effects of certain variables. The fourth question (Explanation Query) asked participants to explain a relationship between two variables in terms of links in the causal model. Table 1 shows the questions from the Abstract condition.

Table 1

Questions and answers for the Abstract conditions of Study 1

Query	Question	Answer
Path	Please list <u>all</u> the ways that variable R could affect variable S	R->S R->M->S R->H->S R->H->A->S R->H->M->S
Cause	Please list <u>all</u> the variables that affect variable A	H, C, R
Effect	Please list <u>all</u> the variables that are affected by variable H	M, S, A
Explanation	Assume that variable H and variable S are found to be positively correlated. Please explain this correlation using the causal model.	- H affects S directly; - H affects M & A which both affect S - R affects H & S directly; - C affects all variables directly or indirectly

The data were analyzed in two ways. If participants listed all of the correct paths/variables and did not list any incorrect paths/variables, the question was scored as correct. A participant's Total Score was the sum of all their correct answers.

However, because each question was composed of multiple answers, it was possible to answer several parts of a question correctly while not answering the entire question correctly. Therefore, to allow partial credit, a second dependent measure was created. This variable was created by summing the number of correct paths or variables each participant listed for a particular question, subtracting the number of incorrect answers, and then dividing that sum by the number of possible correct answers. This created a variable called Proportion-correct Score (PropScore). The maximum score for this variable is 1 (indicating that the answer is entirely correct); the minimum score could be negative.

Results

Participants completed the entire task with a mean response time of 17.6 minutes and a median response time of approximately 11 minutes. Because the presence of outliers strongly affected the mean, median time may be more interpretable. A 2x2 ANOVA was run for Total Time using Visualization and Content. There was not a significant difference between Text and Diagrams, $F(1, 156) = 0.34, p = .558$ or Abstract and Concrete content $F(1, 156) = 0.11, p = .741$. Please see Table 2 for a breakdown of median response times for each question by condition.

Table 2

Median response time (in seconds) by condition for Study 1 Queries

Visualization	Content	Path	Cause	Effect	Expl.	Total Time	N
Text	Abstract	238	91	61	135	785	40
	Concrete	227	86	59	111	655	39
	Marginal Median	231	91	61	123	742	79
Diagram	Abstract	116	79	48	113	564	42
	Concrete	233	79	56	105	735	39
	Marginal Median	175	79	50	110	638	81
Combined	Abstract	159	85	57	127	657	82
	Concrete	230	83	59	108	703	78
	Median	199	84	58	115	668	160

The Total Score variable was analyzed using a 2x2 ANOVA with the following factors: Visualization (Text, Diagram) and Content (Abstract, Concrete). The interaction of Visualization and Content was significant, $F(1,223) = 9.89, p = .009, \eta_p^2 = .04$, and there were significant main effects for both factors. Participants in the Diagram conditions scored higher ($M = 2.11, SD = 1.2$) than participants in the Text conditions ($M = 1.67, SD = 1.3$), $F(1,156) = 4.95, p = .027, \eta_p^2 = .03$. The main effect for Content was in the opposite direction to that expected; participants in the Abstract conditions scored *higher* ($M = 2.20, SD = 1.2$) than participants in the Concrete conditions ($M = 1.58, SD = 1.3$), $F(1,156) = 10.32, p = .002, \eta_p^2 = .06$. These effects may slightly under-estimate the true difference between conditions as 23 of the 79 participants used an external aid such as drawing their own diagram or table (16 of the 23 drew a diagram). Only six participants in the two diagram conditions drew external aids. Please see Table 3 and Figure 3 for further details.

Table 3.

Total score: descriptive statistics by condition.

Visualization	Content	Mean	S.D.	N
Text	Abstract	1.73	1.13	40
	Concrete	1.62	1.41	39
	Marginal Mean	1.67	1.27	79
Diagram	Abstract	2.64	1.01	42
	Concrete	1.54	1.21	39
	Marginal Mean	2.11	1.23	81
Combined	Abstract	2.20	1.16	82
	Concrete	1.58	1.30	78
	Mean	1.89	1.27	160

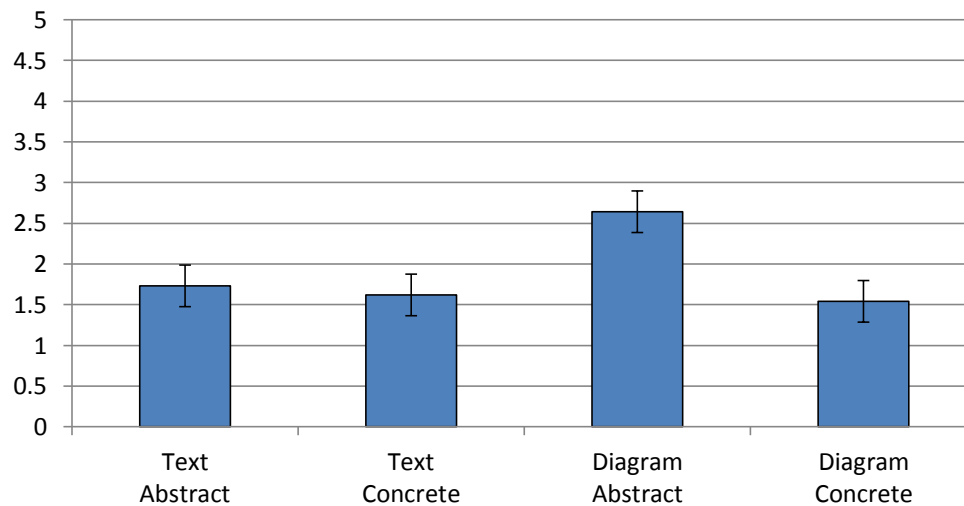


Figure 3. Total Score by condition

The data for each individual Query were also analyzed. For these analyses the proportion-correct score for each question was used. Please see Table 4 for a summary of the average proportion-correct score for each question by condition. The data for each of the Queries were analyzed with a 2x2 ANOVA.

Table 4

Mean proportion-correct score for each Query by condition

Visualization	Content	Path	Cause	Effect	Explanation
Text	Abstract	.56	.67	.50	.26
	Concrete	.55	.55	.48	.31
	Marginal Mean	.55	.61	.49	.28
Diagram	Abstract	.83	.96	.67	.33
	Concrete	.61	.69	.01	.29
	Marginal Mean	.72	.83	.35	.31
Combined	Abstract	.70	.82	.59	.30
	Concrete	.58	.62	.24	.30
	Mean	.64	.72	.42	.30

The Path Query asked the participants to list all the pathways by which one variable could affect another. Both Visualization, $F(1, 156) = 8.72, p = .004, \eta_p^2 = .05$ and Content $F(1, 156) = 4.22, p = .042, \eta_p^2 = .03$, were significant. The Visualization by Content interaction was marginally significant, $F(1, 156) = 3.43, p = .066, \eta_p^2 = .02$.

The Cause Query asked participants to identify all the variables that affected a certain variable. Both Visualization, $F(1, 156) = 7.26, p = .008, \eta_p^2 = .04$ and Content $F(1, 156) = 5.67, p = .019, \eta_p^2 = .04$, were significant. The Visualization by Content interaction was not significant, $F(1, 156) = 0.83, p = .364$.

The Effect Query asked participants to identify all the variables that were affected by another particular variable. Visualization, $F(1, 156) = 1.64, p = .203$, was not significant. There was a main effect for Content however, $F(1, 156) = 8.21, p = .005, \eta_p^2 = .05$. The Visualization by Content interaction was also significant, $F(1, 156) = 7.21, p = .008, \eta_p^2 = .04$.

The Explanation Query asked participants to list potential explanations for a correlation between two variables. No significant differences were found for Visualization, $F(1, 156) = 0.95, p = .332$, Content, $F(1, 156) = 0.01, p = .912$, or the interaction $F(1, 156) = 2.62, p = .108$.

Despite the similarity between the structure of the Cause and Effect Queries, the Effect Query had a lower overall percentage of correct answers. This is especially salient in the Diagram/Concrete condition. The organization of the Queries may have been a contributing factor to this finding. The Cause and Effect Queries were not counterbalanced in their order of appearance for the participants. Consequently it is possible that participants misunderstood the Effect Query (“What variables *are affected by*...”) as asking for the same information as the Cause Query (i.e. “What variables *affect*...”). Evidence that such confusion occurred can be obtained by checking if participants who answered incorrectly were answering the Effect Query as if it were the Cause Query. Analyzing the incorrect answers, it appears this may have happened as 52% of the incorrect answers would have been completely correct had the question been about causes instead of effects. In fact, in the Abstract/Diagram condition, this answer pattern accounted for seven of the eight participants who were incorrect. However, misunderstanding this question was not distributed equally among the conditions. This answer pattern occurred more times in the Concrete/Diagram condition (18) than in all three other conditions combined (17).

Due to the concerns regarding misunderstanding of the Effect Query, the Total Score data were rerun in a 2x2 ANOVA, this time excluding data from the Effect Query. It did not change any of the major findings. The significance of the interaction, $F(1, 156) = 4.78, p = .030, \eta_p^2 = .03$ was slightly less significant. The main effect for Visualization was slightly more significant,

$F(1, 156) = 5.77, p = .013, \eta_p^2 = .04$ and the significance for Content, $F(1, 156) = 4.92, p = .021, \eta_p^2 = .$ was slightly less so. A graph of this analysis can be seen in Figure 4.

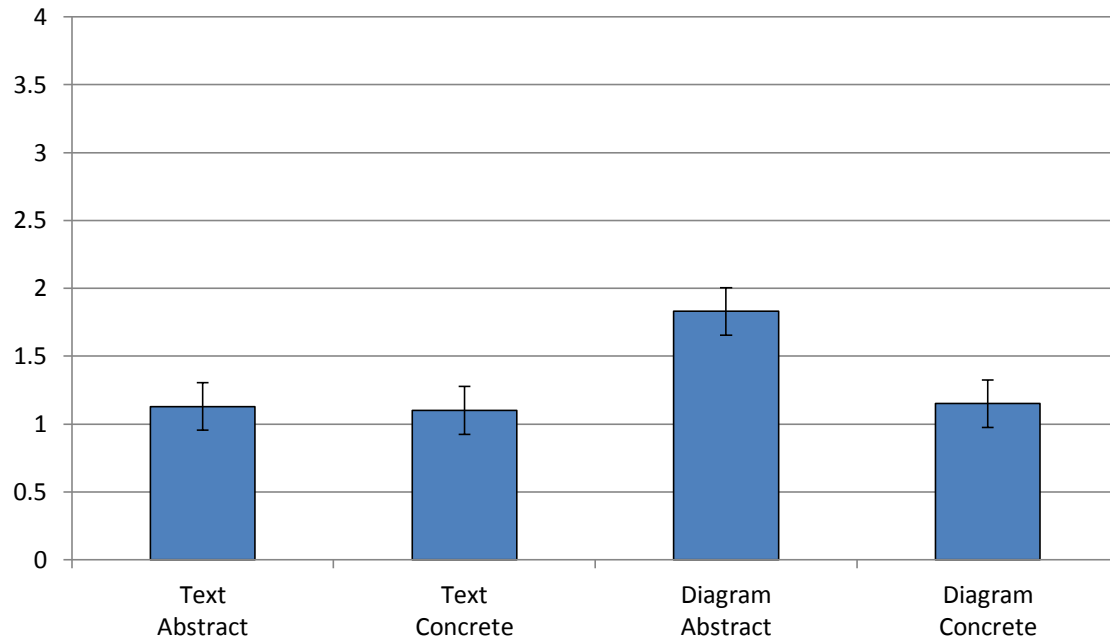


Figure 4. The results of the 2x2 ANOVA with the data from the Effect Query removed.

Discussion

The purpose of this study was twofold: to replicate previous work that showed better performance for tasks completed with a diagrammed causal model over a causal model presented as text and to examine if task performance would be different when reasoning about a model instantiated with real-world content rather than reasoning about a structurally identical model comprised of abstract content.

Previous work on diagrams and causal reasoning has indicated that the use of diagrams resulted in participants working faster and more accurately on performance tasks than when using an identical text version (Corter et al., 2011). Participants in this study who were provided with a diagrammed model had significantly more correct answers when compared to participants who were provided with a text version of the causal model, replicating the previous research.

The second purpose of this study was to examine the effects of problem content on reasoning from a causal model. Causal models generally exist as a tool to aid interpretation of variables that are already placed in a domain context. Placing variables in context might alleviate some of the cognitive effort necessary to interpret the causal model due to previous real-world (or specific domain) experience. Using stimuli from the experiment as an example, it stands to reason that “increased smoking can lead to an increase in premature death” should be easier to remember than “an increase in A leads to an increase in S”. The former is a familiar scenario which may be easier to remember, or may provide details helpful for inference (Baranes, Perry, & Stigler, 1989; Kieras, 1978; Koedinger & Nathan, 2004; Schwanenflugel & Shoben, 1983) while the abstract version holds no inherent significance.

However, the effect was in the reverse direction. Participants scored higher when using abstract models, especially abstract diagrams, than when using models that were composed of real-world variables. Each question that participants answered was designed to assess some particular subskill in interpreting a causal model and each question had several answers. Participants could be incorrect by either providing an incorrect answer, or by failing to provide all the correct answers. The superior performance in the Abstract/Diagram condition was due to those participants identifying more answers than participants in the other three conditions. The number of incorrect answers was minimal and roughly equivalent across all conditions.

The idea that situated context impedes performance is not a new finding. Research finds that context can interfere with reasoning when the task is poorly understood (Cummins, Kintsch, Reusser, & Weimer, 1988; Geary, 1994), when the context is counterintuitive (Easterday, Alevan, & Scheines, 2007; Stroop, 1992), and when the context introduces superfluous information that distracts from the problem (Day, Manlove, & Goldstone, 2011).

Although it is possible that the task was poorly understood in the Concrete condition, the difference in performance was not consistent across conditions, only being found in the Diagram condition. Participants in the Text conditions scored similarly to each other. Additionally, the structure of the questions was identical save for the names of the variables. This is noteworthy because when difficulties arise in using concrete material for mathematical problems, the problem is because the context distracted from understanding the underlying equation. The task was straightforward and not written differently between Abstract and Concrete conditions. It seems unlikely for the task instructions to have been concealed because of distracting contextual information. The relationships between all variables were pilot tested for plausibility so the impediment was probably not due to participant's knowledge running counter to the presented information (Easterday, Aleven, Scheines, & Carver, 2009). It is possible the real-world variables introduced distracting information, although it is unclear what that could have been or how it affected performance.

An unexpected pattern of results for the Effect Query resulted from a programming error. Due to this error, the order of the questions was fixed with the Cause Query always preceding the Effect Query, rather than counterbalancing their order of presentation. These Queries were similar in structure but required reasoning either predictively (from cause to effect) or diagnostically (from effect to cause). The Cause Query asked about the variables that *affected* a certain variable and the Effect Query asked about the variables that *are affected by* a certain variable. It seems that some participants may have interpreted and answered the Effect Query as if it were asking for the same information as the Cause Query. As noted in the Results section, an exploration of the answer patterns indicated this was the case for a large percentage of participants who answered incorrectly.

However, eliminating the Effect query from the analysis did not change the pattern or significance of the results.

One possible explanation for this error in understanding the instructions is that because the Text conditions were more difficult to answer correctly, additional errors were produced that precluded them from being classified in this way. That is to say, this error was defined as participants who would have been correct had the question indeed been reversed. Perhaps other participants were overlooked who made this error but made other errors as well. Because they would not have been correct, even if the question text had been different, they would not have been included in the percentage of participants who fit the pattern of error.

Another possible explanation is that participants in the Diagram conditions had different reasoning systems activated, i.e., System I which involves intuitive, heuristic, automatic reasoning. This system contrasted with System II which involves controlled, analytic, deliberate reasoning (Alter, Oppenheimer, Epley, & Eyre, 2007; Stanovich & West, 2000). The Alter et al. research in particular showed that external cues indicating a difficult problem led people to process information deliberately and analytically whereas cues indicating that a problem is simple led to processing information intuitively and heuristically. Another study showed that when diagrams were used, it triggered higher-order mental processes but often diagrams were not analyzed but instead glossed over (Cromley, Snyder-Hogan, & Luciw-Dubas, 2010). Participants may have considered the Diagram conditions to be simple enough to not activate more deliberate reasoning. Because System I is prone to heuristic judgments, it may be fruitful to examine the structure of causal diagrams to investigate whether any features of the model lead to systematic biases.

Study 2: Connections

Study 1 examined the content of a causal model. Study 2 examined whether changing the diagrammatic conventions of the causal model would lead to systematic biases in reasoning. Specifically, Study 2 examined the type of links between variables and if they change inference performance outcomes between a diagrammed causal model and a text version. Causal models presented as text differ from causal models presented as diagrams by using words instead of arrows to connect variables and by the spatial organization of the content. By replacing the words with arrows and vice versa, the extent to which the use of symbols contributes to participant reasoning can be determined.

This study also sought to address the possible order effects from the presentation of the Cause and Effect Queries in Study 1. Participants were presented with these Queries in random order.

Methods

Participants

A total of 240 participants were recruited from Amazon's Mechanical Turk (MT). Three participants were excluded because they did not complete the task. Another four participants were excluded because they had previously participated in Study 1. Lastly, because the task is relatively contingent upon educational culture, 56 additional participants were not included in the final analysis because they had not completed at least two years of tertiary education. This resulted in 177 remaining participants randomly assigned to four conditions.

The participants were 46% male. Approximately 41% of participants reported their highest level of education attained was an undergraduate degree and 26% reported having attended at least some graduate or professional schooling. About 18% of participants that

reported college majors were from mathematical/engineering/computer science programs. About 60% of participants indicated having taken only one or two classes on statistics and 34% reported never having taken a statistics class. The majority of participants listed English as their first written language (93%). The average age of the participants was 34 years with a range from 18 to 85 years.

Stimuli

The causal model in this study was structurally identical to the one from the previous study. In this version, all the variables are abstract and the manipulation is the type of connection between variables. The connections are either displayed as a word (“affects”) or as an arrow. Please see Figure 5 for the stimuli used in this study.

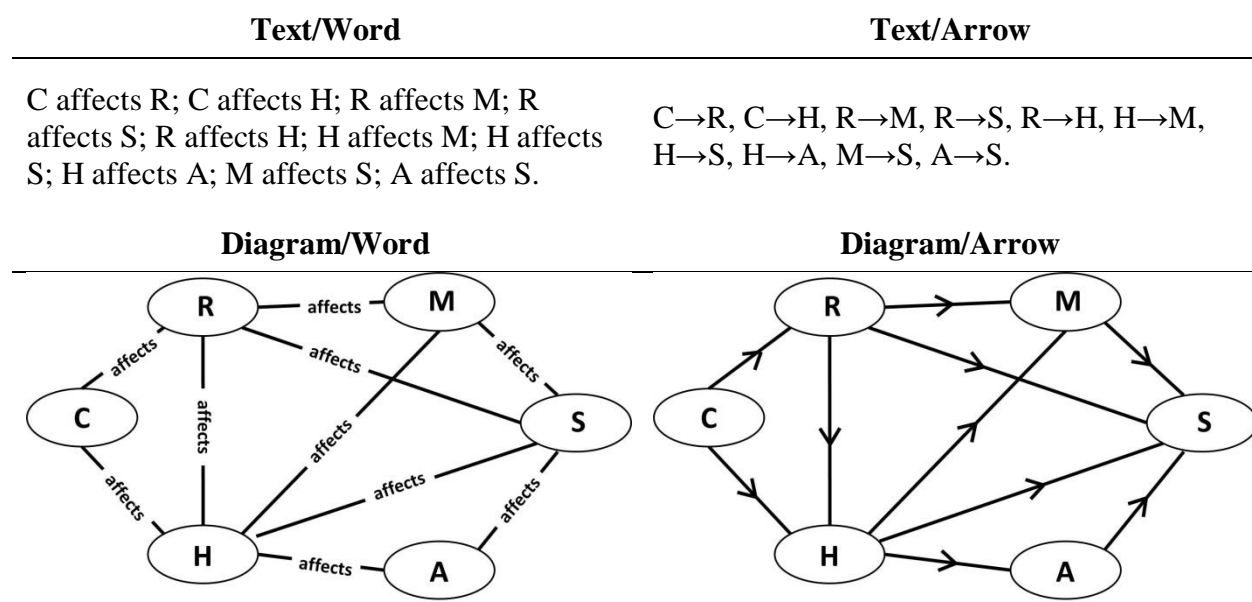


Figure 5. The four conditions for displaying the causal model in Study 2.

Procedure

Our task was hosted online by MT. It was written in HTML and JavaScript. Participants were paid \$1.00 to complete the task. Once participants accepted the task, they were randomly assigned to one of the four conditions and presented with a short introduction. They were

allowed a maximum of one hour from the time they accepted the task to fully complete and submit it. Following is the introduction for all conditions:

In causal modeling of a social science problem, researchers try to specify all the ways in which variables might influence each other.

For example, a researcher might assume that variables X, Y, and Z have the following causal relationships: X affects Y, X affects Z, and Y affects Z.

*In that case, X has a causal influence on Z in two ways. First there is a **direct effect** of X on Z (by assumption). Also there is an **indirect effect** of X on Z, because X is assumed to affect Y and Y is assumed to affect Z.*

To summarize, if you were given the following set of causal assumptions:

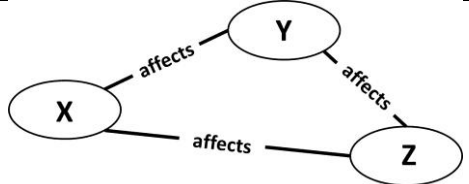
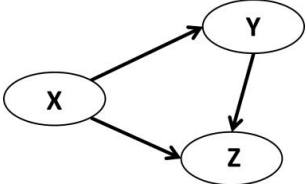
Text/ Word example	Diagram/Word example
<i>X affects Y, X affects Z, and Y affects Z.</i>	 <pre> graph LR X((X)) -- affects --> Y((Y)) X((X)) -- affects --> Z((Z)) Y((Y)) -- affects --> Z((Z)) </pre>
Text/ Arrow example	Diagram/Arrow example
$X \rightarrow Y, X \rightarrow Z, \text{ and } Y \rightarrow Z.$	 <pre> graph LR X((X)) --> Y((Y)) X((X)) --> Z((Z)) Y((Y)) --> Z((Z)) </pre>

Figure 6. The example models for Study 2

The participant was then presented with the causal model in either text or diagram form and using either arrows or words to connect variables (see Figure 6). The instructions continued:

and you were then asked to write the direct and indirect effects of variable X on variable Z, you would write:

X affects Z.

X affects Y which affects Z.

After the introductory example, participants clicked the “proceed” button and saw the following task description and the target causal model:

Now, assume that a researcher makes the following causal assumptions about a particular social science domain where variables C, R, H, M, A, and S are measurable aspects of people.

The researcher makes the following causal assumptions:

The causal model, one of the four versions shown in Figure 5, was presented at this point. With the preceding text and the model visible on the screen, participants were presented with the first of four questions. After they answered that question, they pressed a button and the first question (along with their answer) disappeared and the second question appeared. Meanwhile, the model remained on the screen. This was repeated until the final question which, upon completion, directed participants to a new page asking for demographic information. From this page they submitted their entire task. Participants were not permitted to go back to a previous screen to view or change answers.

Outcome measures

Participants were asked four questions regarding various aspects of the causal models. The questions and answers were identical across conditions. The first question (Path Query) asked participants to identify all direct and indirect effects between two variables. The second and third questions (Cause Query and Effect Query respectively) asked participants to identify causes and effects of certain variables. To control for practice and other order effects, the order in which these two questions appeared was randomized for each participant. The fourth question (Explanation Query) asked participants to explain a relationship between two variables in terms

of links in the causal model. Please see Table 5 for the specific wording of the questions and their respective answers.

The data were analyzed in two ways. If participants listed all of the correct paths/variables and did not list any incorrect paths/variables, the question was scored as correct. A participant's Total Score was the sum of all their correct answers.

Table 5

Study 2 tasks and answers

Query	Question	Answer
Path	Please list all the ways that variable R could affect variable S	R->S R->M->S R->H->S R->H->A->S R->H->M->S
Cause	Please list all the VARIABLES that AFFECT variable A (just name the variables, don't list paths).	H, C, R
Effect	Please list all the VARIABLES that ARE AFFECTED BY variable H (just name the variables, don't list paths).	M, S, A
Explanation	Assume that variable H and variable S are found to be positively correlated. Please explain this correlation using the causal model.	- H affects S directly; - H affects M & A which both affect S - R affects H & S directly; - C affects all variables directly or indirectly

However, because each question was composed of multiple answers, it was possible to answer several parts of a question correctly while not answering the entire question correctly. Therefore, to allow partial credit, a second dependent measure was created. This variable was created by summing the number of correct paths or variables each participant listed for a particular question, subtracting the number of incorrect answers, and then dividing that sum by the number of possible correct answers. This created a variable called Proportion-correct Score

(PropScore). The maximum score for this variable is 1 (indicating that the answer is entirely correct); the minimum score could be negative.

Results

The mean time participants took to complete the task was 10.5 minutes. However, this included several outliers that took over 45 minutes. The median time participants took to complete the task was 8.5 minutes. Time to complete the task was analyzed using a 2x2 ANOVA and found a significant effect for Visualization, $F(1,173) = 5.92, p = .016, \eta_p^2 = .03$. There was not a significant effect for Connection, $F(1,173) = 1.64, p = .202$. A breakdown of median response times for each question by condition can be found in Table 6.

Table 6

Median response time (in seconds) by condition for Study 2 Queries

Visualization	Connection	Path	Cause	Effect	Expl.	Total Time	N
Text	Word	209	79	64	130	567	47
	Arrow	173	79	63	96	554	41
	Marginal Median	195	79	64	113	566	88
Diagram	Word	142	43	26	108	482	43
	Arrow	106	41	30	83	345	46
	Marginal Median	124	42	29	94	414	89
Combined	Word	165	61	40	115	552	90
	Arrow	126	52	38	87	448	87
	Median	146	52	38	100	507	177

The Total Score variable was analyzed using a 2x2 ANOVA with the following factors: Visualization (Text, Diagram) and Connection (Word, Arrow). Based on previous studies, a significant effect for Visualization was predicted although not in the direction that occurred. The

interaction of Visualization and Connection produced a significant effect, $F(1,173) = 65.94, p < .001, \eta_p^2 = .38$. Participants in the Diagram conditions scored *lower* ($M = 1.51, SD = 1.5$) than participants in the Text conditions ($M = 1.98, SD = 1.1$), $F(1,173) = 11.47, p < .001, \eta_p^2 = .07$. Participants in the Arrow conditions scored significantly higher ($M = 2.28, SD = 1.1$) than participants in the Word conditions ($M = 1.22, SD = 1.3$), $F(1,173) = 51.57, p < .001, \eta_p^2 = .30$. This is due to very poor performance in the Diagram/Word condition. Nineteen participants drew their own diagram or table in the Text conditions. Only two participants, both in the Diagram/Word condition, drew their own diagrams. Please see Table 7 and Figure 7 for further details.

Table 7

Descriptive statistics for all four conditions

Visualization	Connection	Total Score Mean	S.D.	N
Text	Word	2.04	1.0	47
	Arrow	1.90	1.1	41
	Marginal Mean	1.98	1.1	88
Diagram	Word	0.33	0.8	43
	Arrow	2.61	1.0	46
	Marginal Mean	1.51	1.5	89
Combined	Word	1.22	1.3	90
	Arrow	2.28	1.1	87
	Mean	1.74	1.3	177

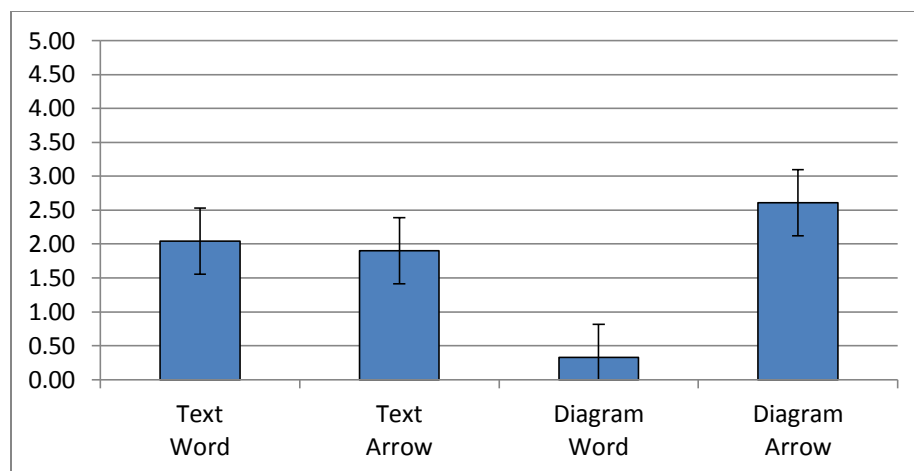


Figure 7. Total score by condition

The results indicate that something in the Diagram/Word condition impeded correct task completion. The proportional-correct score was analyzed for each Query using a 2x2 ANOVA. The poor performance for participants in the Diagram/Word condition was not limited to any particular question but rather was consistent throughout the entire experiment. Please see Table 8 for a summary of the answers for each Query broken down by condition.

Table 8

Mean proportion-correct score for Query by condition

Visualization	Connection	Path	Cause	Effect	Explanation
Text	Word	.73	.81	.83	.28
	Arrow	.68	.78	.85	.29
	Marginal Mean	.71	.80	.84	.28
Diagram	Word	.27	.19	.33	.16
	Arrow	.84	.86	.89	.29
	Marginal Mean	.57	.54	.62	.23
Combined	Word	.51	.51	.59	.22
	Arrow	.77	.82	.87	.29
	Mean	.64	.66	.73	.25

Path Query asked the participants to list all the pathways by which one variable could affect another. Both Visualization, $F(1, 173) = 9.12, p = .003, \eta_p^2 = .05$ and Connection $F(1, 173) = 28.38, p < .001, \eta_p^2 = .14$, were significant. The Visualization by Connection interaction was also significant, $F(1, 173) = 41.488, p < .001, \eta_p^2 = .19$.

The Cause Query asked participants to identify all the variables that affected a certain variable. Both Visualization, $F(1, 173) = 19.69, p < .001, \eta_p^2 = .10$ and Connection $F(1, 173) = 27.06, p < .001, \eta_p^2 = .14$, were significant. The Visualization by Connection interaction was also significant, $F(1, 173) = 32.07, p < .001, \eta_p^2 = .16$.

The Effect Query asked participants to identify all the variables that were affected by another particular variable. Visualization, $F(1, 173) = 10.59, p = .01, \eta_p^2 = .06$, was significant. There was also a main effect for Connection however, $F(1, 173) = 16.92, p < .001, \eta_p^2 = .09$. The Visualization by Connection interaction was significant, $F(1, 173) = 14.29, p < .001, \eta_p^2 = .08$.

The Explanation Query asked participants to offer potential explanations for a correlation between two variables. Both Visualization, $F(1, 173) = 4.42, p = .037, \eta_p^2 = .03$ and Connection $F(1, 173) = 6.40, p = .012, \eta_p^2 = .04$, were significant. The Visualization by Connection interaction was also significant, $F(1, 173) = 4.65, p = .032, \eta_p^2 = .03$.

The results show that the participants in the Diagram/Word condition did worse on every component in the task. As explained in the section on outcome measures, participants received a score for each question based on two criteria: number of correct answers and number of incorrect answers. It is possible for someone to have gotten all the correct answers, but have a diminished

score because of incorrect answers as well. As Table 9 shows, participants in the Diagram/Word condition did similarly on identifying correct answers but listed more incorrect answers.

Table 9

Average of correct and incorrect answers by condition.

	Path		Cause		Effect	
	Corr.	Incorr.	Corr.	Incorr.	Corr.	Incorr.
Text/Word	4.0	0.4	2.7	0.3	2.8	0.3
Text/Arrow	3.7	0.3	2.7	0.3	2.8	0.2
Diagram/Word	4.0	2.6	2.4	1.8	2.9	1.9
Diagram/Arrow	4.5	0.2	2.8	0.3	2.9	0.2

Discussion

The purpose of this study was twofold: to replicate previous work on the presentation of causal models and to examine the effect link representation has on one's ability to interpret a causal model. This study failed to fully replicate previous findings regarding higher performance for participants with diagrams. The diagram using traditional arrows scored higher than the other conditions. However, the diagram using words scored so much lower than the other conditions that the analysis returned a main effect that text outperformed diagrams. As suggested in Study 1, perhaps reasoning with a diagram activates the intuitive system of reasoning. Some aspect of the diagram that would not be an obstacle in analytic reasoning might create a reasoning bias with intuitive and heuristic reasoning.

Two important differences exist between a diagrammed representation of a causal model and a textual version of the same model—the spatial array and the links between variables. A link in a causal model indicates the direct relationship between variables and the direction of their causal relationship and is typically represented as a word or an arrow. Both arrows and

words share asymmetric properties. An arrow is by nature directional and when comprised with only one arrowhead, indicative of an asymmetrical relationship. In the current study, the word “affects” indicated the causal link between variables. As a word in the English language, it is read left-to-right and as a verb, implies that the subject acts upon the predicate, e.g. C affects R. The word “affects” would not be read to mean the opposite. If one were to keep the same sentence structure but reverse the causal direction so that R is the agent that acts upon C, the verb “affects” would have to be replaced by the verb phrase “is affected by”. Given the similarity in purpose between the word and the arrow, one could reasonably assume that by embedding words into the lines of the diagram, a word could be seen as equivalent to an arrow for indicating causal direction.

The primary explanation for the poor performance in the Diagram/Word condition was because the participants had a higher number of errors than participants in all other conditions combined. The average number of errors in the other conditions (≈ 0.3) is consistent with the average number of errors in Study 1. The majority of incorrect answers in the Diagram/Word condition appeared to stem from misinterpreting the meaning of the link between variables. The participants answered the questions as if the link were correlational rather than causal. For example, when presented with the diagram showing variable C affects variable R, participants interpreted the inverse as well, that R affects C. In the path Query, where the answer was to write the five correct paths, participants wrote on average 2.5 incorrect paths that were almost exclusively incorrect because they reversed the causal direction of the variables (i.e. the answers would have been correct if the arrows had indeed been bidirectional). In the Cause and Effect Queries, if participants continued to interpret the causal direction as bidirectional, then there

would have been an answer pattern of three correct variables and 2 incorrect ones. As shown by Table 9, this is indeed what appears to have happened.

By replacing the arrow with the word “affects”, participants were seen to have been influenced to think of the causal relationships as bidirectional. It impeded their interpretation of the model not through complexity but rather because the presence of a word instead of an arrowhead activated a bias that ignored the literal definition of the word.

There are several possible explanations of this effect. The most apparent is that the word was placed in the middle of the arc connecting variables, and not at the end, as it would have been if it were still a typical arrow. Although the arrowheads were also placed in the middle of the arc, the asymmetric nature of the arrowhead symbol may be more salient. Another possible explanation is the nature of the word itself. It is unclear whether this effect would hold if we used a different, less ambiguous word e.g. C causes R. Also, perhaps the effect would be mitigated with the addition of more specific instructions. Our current instructions imply that the word “affects” is unidirectional in its causal implications. The inclusion of an additional instruction explicitly stating its unidirectionality could aid participants, as could a simple test to ensure participants were clear on how the model worked.

Participants in the other conditions did not need such explicit instruction and they did roughly the same in the Text condition with either a word link or an arrow link. In a sentence, an arrow seems to be able to replace text with less difficulty than text replacing an arrow in a diagram. This may be another example of the difference between intuitive and analytical reasoning. As the diagram may activate the intuitive reasoning, perhaps text activated analytical reasoning.

Study 3: Replication of Connections Study

Study 3 sought to examine whether the results of Study 2 were an actual effect of inhibitory performance caused by using words instead of arrowheads, or whether they were due to poor task construction. Two areas that presented alternative explanations were the instructions and the location of the words on the arcs connecting variables. Regarding the instructions, participants were required to take a short qualification exam which they had to pass in order to be included in this study. And regarding the placement of the words in the diagram, they were moved to the end of the arc to more closely mimic typical arrowheads. The arrowheads were also moved to their normal position.

Methods

Participants

A total of 234 participants were recruited from Amazon's Mechanical Turk (MT). Four participants were excluded because they did not complete the task. Another 33 participants were excluded because they had previously participated in another study. As before, because the task is relatively contingent upon educational culture, 68 additional participants were not included in the final analysis because they had not completed at least two years of tertiary education. Lastly, 25 participants were excluded because they did not pass the qualification questions. This resulted in 104 remaining participants randomly assigned to four conditions.

The participants were 61% male. Approximately 46% of participants reported their highest level of education attained was an undergraduate degree and 22% reported having attended at least some graduate or professional schooling. About 32% of participants that reported college majors were from mathematical/engineering/computer science programs. About 66% of participants indicated having taken only one or two classes on statistics and 30%

reported never having taken a statistics class. The majority of participants listed English as their first written language (94%). The average age of the participants was 31 years with a range from 18 to 69 years.

Stimuli

The causal model in this study was the same as the one from the previous study with the exception that in the Diagram conditions, the arrowheads and the word-arrowheads were moved to the end of the connecting arc. Please see Figure 8 for the stimuli used in this study.

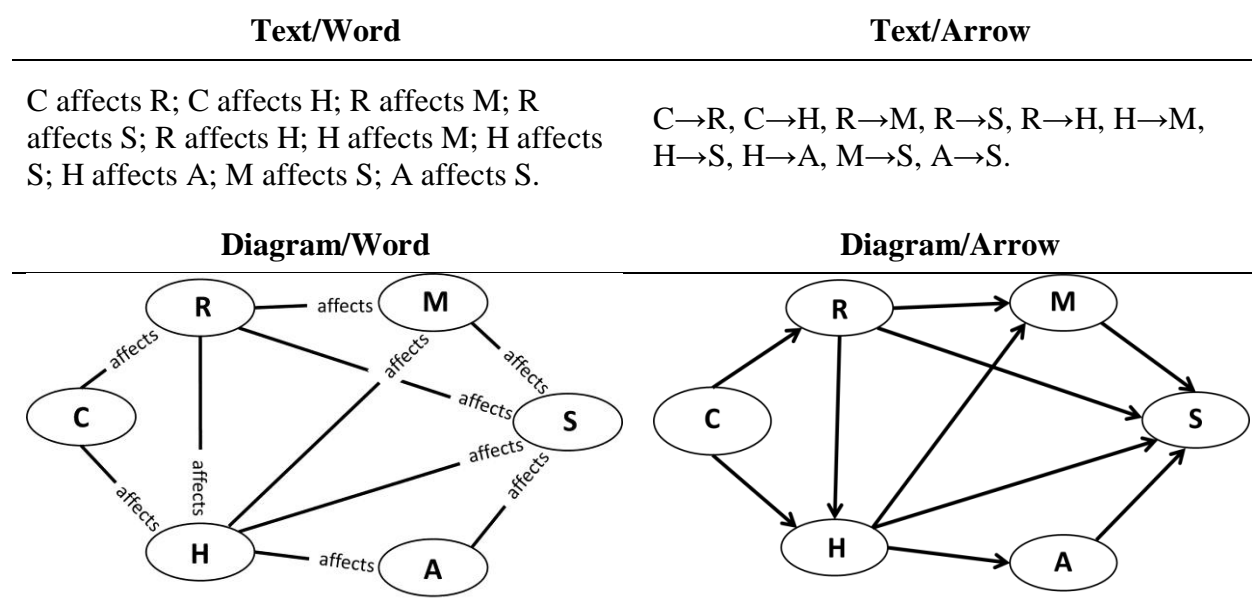


Figure 8. The four conditions for displaying the causal model in Study 3.

Procedure

Our task was hosted online by MT. It was written in HTML and JavaScript. Participants were paid \$1.00 to complete the task. Once participants accepted the task, they were randomly assigned to one of the four conditions and presented with a short introduction. They were allowed a maximum of one hour from the time they accepted the task to fully complete and submit it. Following is the introduction for all conditions:

In causal modeling of a social science problem, researchers try to specify all the ways in which variables might influence each other.

For example, a researcher might assume that variables X, Y, and Z have the following causal relationships: X affects Y, X affects Z, and Y affects Z.

*In that case, X has a causal influence on Z in two ways. First there is a **direct effect** of X on Z (by assumption). Also there is an **indirect effect** of X on Z, because X is assumed to affect Y and Y is assumed to affect Z.*

To summarize, if you were given the following set of causal assumptions:

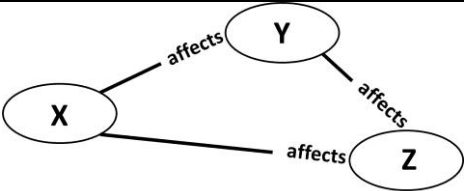
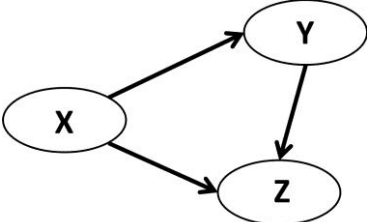
Text/ Word example	Diagram/Word example
<i>X affects Y, X affects Z, and Y affects Z.</i>	
Text/ Arrow example	Diagram/Arrow example
$X \rightarrow Y, X \rightarrow Z, \text{ and } Y \rightarrow Z.$	

Figure 9. The example models for Study 3

The participant was then presented with the causal model in either text or diagram form and using either arrows or words to connect variables (see Figure 9). The instructions continued:

and you were then asked to write the direct and indirect effects of variable X on variable Z, you would write:

X affects Z.

X affects Y which affects Z.

Here the new qualification questions were asked. The first was whether Y affected X and they could check either “yes”, “no”, or “can’t tell”. The correct answer was “no”. The next

question asked whether Y affected Z with the same answer options. This answer was “yes”.

After the introductory example and qualification test, participants clicked the “proceed” button and saw the following task description and the target causal model:

Now, assume that a researcher makes the following causal assumptions about a particular social science domain where variables C, R, H, M, A, and S are measurable aspects of people.

The researcher makes the following causal assumptions:

The causal model, one of the four versions shown in Figure 8, was presented at this point. With the preceding text and the model visible on the screen, participants were presented with the first of four questions. After they answered that question, they pressed a button and the first question (along with their answer) disappeared and the second question appeared. Meanwhile, the model remained on the screen. This was repeated until the final question which, upon completion, directed participants to a new page asking for demographic information. From this page they submitted their entire task. Participants were not permitted to go back to a previous screen to view or change answers.

Outcome measures

Participants were asked three questions regarding various aspects of the causal models. The questions and answers were identical across conditions. The first question (Path Query) asked participants to identify all direct and indirect effects between two variables. The second and third questions (Cause Query and Effect Query respectively) asked participants to identify causes and effects of certain variables. To control for practice and other order effects, the order in which these two questions appeared was randomized for each participant. Please see Table 10 for the specific wording of the questions and their respective answers.

The data were analyzed in two ways. If participants listed all of the correct paths/variables and did not list any incorrect paths/variables, the question was scored as correct.

A participant's Total Score was the sum of all their correct answers.

Table 10

Study 3 tasks and answers

Query	Question	Answer
Path	Please list <u>all</u> the ways that variable R could affect variable S	R->S R->M->S R->H->S R->H->A->S R->H->M->S
Cause	Please list all the VARIABLES that AFFECT variable A (just name the variables that are <u>causes</u> of A, don't list paths).	H, C, R
Effect	Please list all the VARIABLES that variable H AFFECTS (just name the variables that are <u>effects</u> of H, don't list paths).	M, S, A

However, because each question was composed of multiple answers, it was possible to answer several parts of a question correctly while not answering the entire question correctly. Therefore, to allow partial credit, a second dependent measure was created. This variable was created by summing the number of correct paths or variables each participant listed for a particular question, subtracting the number of incorrect answers, and then dividing that sum by the number of possible correct answers. This created a variable called Proportion-correct Score (PropScore). The maximum score for this variable is 1 (indicating that the answer is entirely correct); the minimum score could be negative.

Results

The mean time participants took to complete the task was 9 minutes. However, this included several outliers that took over 30 minutes. The median time participants took to complete the task was 5.5 minutes. The time it took to complete the task was analyzed using a 2x2 ANOVA. No significant effect was found for Visualization, $F(1,100) = 0.002$, $p = .969$, or for Connection, $F(1,100) = 2.72$, $p = .102$. A breakdown of median response times for each question by condition can be found in Table 11.

Table 11

Median response time (in seconds) by condition for Study 3 Queries

Visualization	Connection	Path	Cause	Effect	Total Time	N
Text	Word	145	66	66	397	18
	Arrow	129	51	41	338	33
	Marginal Median	134	60	55	372	51
Diagram	Word	118	42	35	310	20
	Arrow	94	39	29	286	32
	Marginal Median	108	40	30	288	52
Combined	Word	126	51	50	373	38
	Arrow	113	45	32	307	65
	Median	121	49	36	336	103

The Total Score variable was analyzed using a 2x2 ANOVA with the following factors: Visualization (Text, Diagram) and Connection (Word, Arrow). The interaction of Visualization and Connection produced a significant difference, $F(1,100) = 3.84$, $p = .053$, $\eta_p^2 = .04$. Participants in the Diagram/Arrow condition scored on average a full point higher than the other three conditions. There was also a significant effect for Visualization, $F(1,100) = 4.31$, $p = .041$,

$\eta_p^2 = .04$, with Diagrams outperforming Text. There was not a significant difference for Connection, $F(1,100) = 0.59$, $p = .444$. The difference between Text and Diagram conditions may be under-estimated as 24 participants in the Text conditions drew their own diagrams and six drew tables. Only nine participants over both Diagram conditions drew external aids (seven drew tables). Please see Table 12 and Figure 10 for further details.

Table 12

Descriptive statistics for all four conditions

Visualization	Connection	Total Score Mean	S.D.	N
Text	Word	1.58	0.7	19
	Arrow	1.36	0.9	33
	Marginal Mean	1.44	0.8	52
Diagram	Word	1.60	1.2	20
	Arrow	2.09	0.8	32
	Marginal Mean	1.90	1.0	52
Combined	Word	1.59	1.0	39
	Arrow	1.72	0.9	65
	Mean	1.67	0.9	104

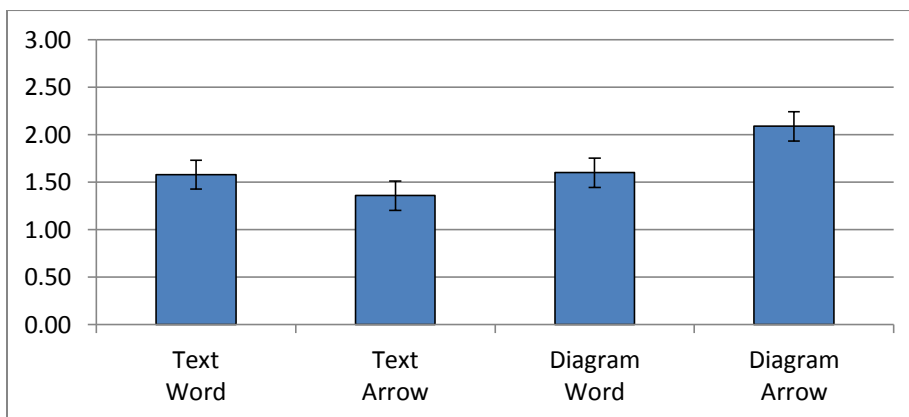


Figure 10. Total score by condition

The proportional-correct score was analyzed for each individual Query using a 2x2 ANOVA. Please see Table 13 for a summary of the answers for each Query broken down by condition.

Table 13

Mean proportion-correct score for Query by condition

Visualization	Connection	Path	Cause	Effect
Text	Word	.72	.74	.88
	Arrow	.61	.69	.66
	Marginal Mean	.65	.71	.74
Diagram	Word	.62	.75	.60
	Arrow	.81	.74	.93
	Marginal Mean	.74	.74	.80
Combined	Word	.67	.74	.74
	Arrow	.71	.71	.79
	Mean	.69	.72	.77

Path Query asked the participants to list all the pathways by which one variable could affect another. The Visualization by Connection interaction was significant, $F(1, 100) = 6.81, p = .01, \eta_p^2 = .06$. Neither Visualization, $F(1, 100) = 0.85, p = .359$, nor Connection $F(1, 100) = 0.61, p = .436$, was significant.

The Cause Query asked participants to identify all the variables that affected a certain variable. There was no significance found for Visualization, $F(1, 100) = 0.15, p = .697$, Connection $F(1, 100) = 0.13, p = .721$, or the Visualization by Connection interaction, $F(1, 100) = .06, p = .815$.

The Effect Query asked participants to identify all the variables that were affected by another particular variable. The Visualization by Connection interaction was significant, $F(1, 100) = 7.34, p = .008, \eta_p^2 = .07$. Neither Visualization, $F(1, 100) = 0.001, p = .974$, nor Connection were significant, $F(1, 100) = 0.28, p = .600$.

Discussion

Study 3 was designed to control for some of the alternative explanations that arose in Study 2. In Study 2, the results indicated that by replacing arrowheads with words, participants were greatly affected by being more prone to error. The particular error observed in this condition was that participants seemed to interpret the word placed in the middle of the arc as being bidirectional. That is to say, A affects B was consistently reported as B also affecting A. One possibility for this error was that the instructions were simply unclear as the diagram was by no means a typical causal representation. Another possibility was that by placing the word in the middle of the arc connecting variables, the positioning also cued participants to interpret this atypical representation as being bidirectional. Study 3 moved the placement and created a qualification test to ensure that participants knew these representations were not bidirectional before proceeding with the task.

Participants in the Diagram/Word condition did much better this time, scoring on par with participants in the Text conditions. However, the Diagram/Arrow representation still outperformed all other conditions. And as before, the reason appears to be because of errors made in the diagram/Word condition. The most common error was again misreading the links as being bidirectional, despite participants having already passed the test indicating they understood the links were not bidirectional. The number of errors was less than in the previous study, indicating that perhaps the qualification test did help to decrease the error.

These results lend support to the idea of the last two studies that diagrams may activate System I reasoning which causes participants faced with an unfamiliar situation to erroneously rely on intuitive interpretations of a causal model. These two studies indicate that further exploration of appropriate model construction is needed in order to prevent other errors that may arise in more typical uses of causal models. It would be interesting to examine whether this effect would hold up if participants were not using abstract variables

Study 4: Asymmetries in reasoning about causes and effects

The primary goal of Study 4 was to examine the possibility of asymmetries between the ease or accuracy of reasoning from cause to effect (predictive reasoning) and from effect to cause (diagnostic reasoning). Previous work in this area focused on these types of reasoning in the context of a mental simulation (Hagmayer & Waldmann, 2000; Hegarty, 1992) or in a word problem (Fernbach & Sloman, 2009; Fernbach, Darlow, & Sloman, 2011; Tversky & Kahneman, 1980). Although predictive reasoning was often favored in these circumstances, it is unclear whether the effect was due to a preference for predictive reasoning or because some aspect of the problem itself created the bias. Participants in this study were given an explicit but abstract causal model to control for any context effects brought about by prior knowledge.

By taking a causal model and asking about the causes of a certain variable, and then reversing the model and asking about the effects of the same variable, this experiment examined the asymmetry hypothesis. Two different groups of participants were asked to list either all the causes or all the effects of a given variable. In this situation, the models given to the two groups were designed so that the correct answers would be identical, even though the questions would be framed either from a cause perspective or an effect perspective. That is to say, if the causes of variable A in the first model were variables X, Y, and Z, then the effects of variable A in the second model were also variables X, Y, and Z.

A secondary purpose for this study was to further examine findings from previous studies regarding the circumstances in which diagrams improve reasoning compared to text. Studies 1 and 2 offered evidence that diagrammed causal models may activate intuitive reasoning and causal models presented as text may activate more analytical reasoning. Study 4 examined

whether this difference in reasoning created a performance difference between predictive and diagnostic reasoning.

The last objective of this study was to examine the ability of people to discover alternative explanations for variable relationships. The previous two studies asked participants to explain the relationship between two variables that were related through direct effect, indirect effects, and a common cause. Participants had limited success in this particular task overall, but were especially poor at discovering the common cause explanation. However, the answers to those questions were relatively complex, involving three types of explanations. Thus it may have been excessive to expect so much from people without training in causal modeling or statistics. In particular, the presence of more readily identifiable alternative explanations (the direct and indirect effect links) may create an effect similar to phonological blocking (Woodworth, 1929) where an initial explanation can prevent access to an alternative (in this case, the alternative being the common cause leading to a spurious correlation). Therefore, the previous studies were unable to determine whether the idea of spurious correlation was too difficult to discover or whether it was simply blocked or obscured. By reducing the complexity of the task, it was a better assessment of whether the concept of a spurious correlation was discoverable within the context of a causal model. Including one alternative explanation (a direct effect link) in one of the models allowed for testing the hypothesis that distractor answers had previously interfered with the discovery of spurious correlations in previous studies.

Methods

Participants

We recruited 240 participants from Amazon's Mechanical Turk (MT) for this study. The experiment was only available to participants who had an American IP address. Four

participants were excluded for not completing the task. Another 21 participants were excluded for having completed one of the previous studies. This resulted in 215 participants randomly assigned to four conditions.

The participants were 60% male. English was the native language for 99% of participants. The mean age was 31 years old ($SD = 11$) and ranged from 18 to 87 years. About 86% of participants reported having attended at least some college with 17% having attended at least some graduate school. Of those who attended college, 34% (31% overall) majored in mathematics, engineering, computer science, or information systems. However, only 5% of all participants reported ever having taken more than two courses in statistics. Over half (55%) reported never having taken a statistics course.

Stimuli

The causal model in this study was structurally identical to the one from the previous studies. While some participants saw this base model (Model A), others saw its inverse (Model B). The inverse model had all causal link directions reversed, and was reflected around the vertical axis so that the general direction of causal flow remained from left to right. The models were also presented as either Text or as a Diagram. Please see Figure 11 for the stimuli used in this study.

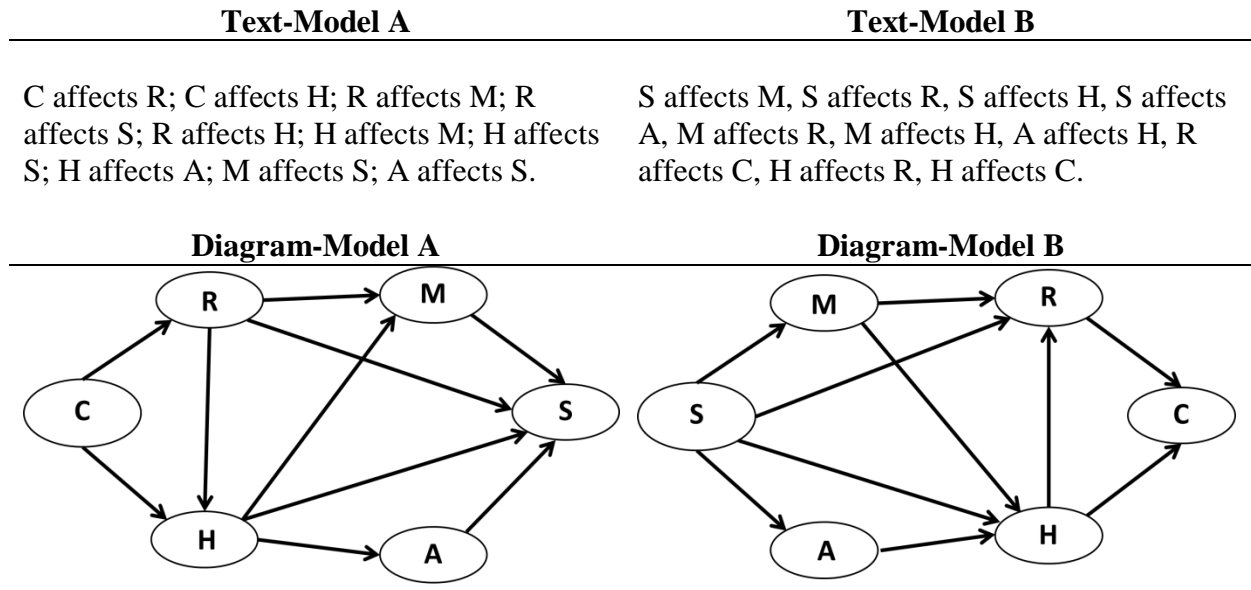


Figure 11. The four conditions for displaying the causal model in Study 4.

Procedure

Our task was hosted online by MT. It was written in HTML and JavaScript. Participants were paid \$1.00 to complete the task. Once participants accepted the task, they were randomly assigned to one of the four conditions and presented with a short introduction. They were allowed a maximum of one hour from the time they accepted the task to fully complete and submit it. Following is the introduction for all conditions:

A primary goal of science is to uncover the causes of the phenomena of interest. Because many phenomena have multiple causes, some direct and some indirect, uncovering causal relationships can be complicated. For example, food consumption is a direct cause of obesity. The number of hours of TV watched per week is an indirect cause because greater TV time reduces energy expenditure. After preliminary research, scientists often model the causal relations they have found, and use the model to come to conclusions.

Consider a simple example, represented by the information below:

X affects Y, X affects Z, and Y affects Z.

*In this case, X has a causal influence on Z in two ways. First there is a **direct effect** of X on Z (by assumption). Also there is an **indirect effect** of X on Z, because X is assumed to affect Y and Y is assumed to affect Z. To summarize, if you were given this information*

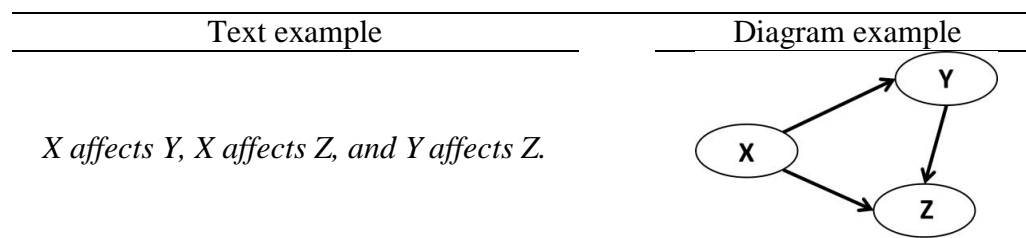


Figure 12. The example models for Study 4

The participant was then presented with the example causal model. Depending on the condition to which they had been assigned, it was either in text or diagram form (see Figure 12).

The instructions continued:

and asked to list all the ways that variable X causally affects variable Z, you would write:

X affects Z.

X affects Y which affects Z.

After the introductory example, participants clicked the “proceed” button and saw the following task description and the target causal model:

Assume that a researcher makes causal assumptions about a particular social science domain where variables C, R, H, M, A, and S are measurable aspects of people.

The researcher makes the following causal assumptions:

The causal model, in one of the four versions shown in Figure 11, was presented at this point. With the preceding text and the model visible on the screen, participants were presented with the first of four questions. After they answered that question, they pressed a button and the first question (along with their answer) disappeared and the second question appeared.

Meanwhile, the model remained on the screen. This was repeated until the final question which, upon completion, directed participants to a new page asking for demographic information. From this page they submitted their entire task. Participants were not permitted to go back to a previous screen to view or change answers.

Outcome measures

Answers were collected from the four questions presented to the participants. The first question (Path Query) asked participants to identify all direct and indirect effects between two variables. The second and third questions (Cause Query and Effect Query respectively) asked participants to identify causes and effects of certain variables. To control for practice and other order effects, the order in which these two questions appeared was randomized for each participant. The fourth question (Explanation Query) asked participants to explain a relationship between two variables in terms of links in the causal model. Please see Table 14 for the specific wording of the questions and their respective answers.

We analyzed the data in two ways for the first three questions. If participants listed all of the correct paths/variables and did not list any incorrect paths/variables, the question was scored as correct. A participant's Total Score was the sum of all their correct answers.

However, because each question was composed of multiple answers, it was possible to answer several parts of a question correctly while not answering the entire question correctly. Therefore, to allow partial credit, a second dependent measure was created. This variable was created by summing the number of correct paths or variables each participant listed for a particular question, subtracting the number of incorrect answers, and then dividing that sum by the number of possible correct answers. This created a variable called Proportion-correct Score

(PropScore). The maximum score for this variable is 1 (indicating that the answer is entirely correct); the minimum score could be negative.

Table 14

Study 4 tasks and answers

Query	Model A		Model B	
	Question Text	Answer	Question Text	Answer
Path	Please list all the ways that variable R could affect variable S.	R->S R->M->S R->H->S R->H->A->S R->H->M->S	Please list all the ways that variable S could affect variable R.	S->R S->M->R S->H->R S->A->H->R S->M->H->R
Cause	Please list all the VARIABLES that AFFECT variable M (just name the variables, don't list paths).	C, H, R	Please list all the VARIABLES that AFFECT variable R (just name the variables, don't list paths).	H, M, A, S
Effect	Please list all the VARIABLES that ARE AFFECTED BY variable R (just name the variables, don't list paths).	H, M, A, S	Please list all the VARIABLES that ARE AFFECTED BY variable M (just name the variables, don't list paths).	C, H, R
Explanation	Assume that variable R and variable H are found to be positively correlated. Please explain this correlation using the causal model.	Both variables affected by C; $R \rightarrow H$	Assume that variable M and variable A are found to be positively correlated. Please explain this correlation using the causal model.	Both variables affected by S

Answers to the Explanation Query were scored somewhat differently. Participants were scored as correct or incorrect based only on whether they identified the spurious correlation.

Results

Participants completed the entire task with a mean response time of 10 minutes and a median response time of 7.5 minutes. Because the presence of outliers strongly affected the mean, median time may more interpretable. Time to complete the task was analyzed using a 2x2 ANOVA and found a significant effect for the interaction between Visualization and Model, $F(1,210) = 4.52, p = .035, \eta_p^2 = .02$. There was also a significant effect for Visualization, $F(1,210) = 12.55, p < .001, \eta_p^2 = .06$. A breakdown of median response times for each question by condition can be found in Table 15.

Table 15

Median response time (in seconds) by condition for Study 4 Queries

Visualization	Model	Path	Cause	Effect	Expl.	Total Time	N
Text	Model A	177	66	44	103	428	57
	Model B	154	54	40	126	460	45
	Marginal Median	164	61	44	112	432	102
Diagram	Model A	118	34	25	70	271	60
	Model B	103	30	20	81	254	52
	Marginal Median	104	33	22	76	263	112
Combined	Model A	151	51	33	88	386	117
	Model B	115	40	31	96	315	97
	Median	134	45	32	94	339	214

Text vs. Diagram

The Total Score variable was analyzed using a 2x2 ANOVA with the following factors: Visualization (Text, Diagram) and Model (Model A, Model B). The participants in the Diagram conditions scored higher ($M = 2.01, SD = 1.0$) than in the Text conditions ($M = 1.49, SD = 1.1$),

$F(1,211) = 16.988, p < .001, \eta_p^2 = .08$. Participants presented with Model A also scored higher ($M = 1.97, SD = 1.0$) than those presented with Model B ($M = 1.51, SD = 1.0$), $F(1,211) = 12.846, p < .001, \eta_p^2 = .06$. The Visualization*Model interaction was marginally significant, $F(1,211) = 2.722, p = .093, \eta_p^2 = .01$. The difference between Diagrams and text may be slight under-estimated as six participants drew their own diagrams and nine drew their own tables. In the Diagram conditions, one person drew another diagram and two drew tables. Please see Table 16 and Figure 13 for further details.

Table 16

Descriptive statistics for all four conditions

Visualization	Model	Total Score Mean	S.D.	N
Text	Model A	1.81	1.0	57
	Model B	1.09	1.0	45
	Marginal Mean	1.49	1.1	102
Diagram	Model A	2.13	0.9	60
	Model B	1.87	1.0	53
	Marginal Mean	2.01	0.9	113
Combined	Model A	1.97	1.0	117
	Model B	1.51	1.1	98
	Mean	1.76	1.0	215

Total Score is computed as the sum of the proportional correct scores for Path, Cause, and Effect Queries and omitting the Explanation Query score.

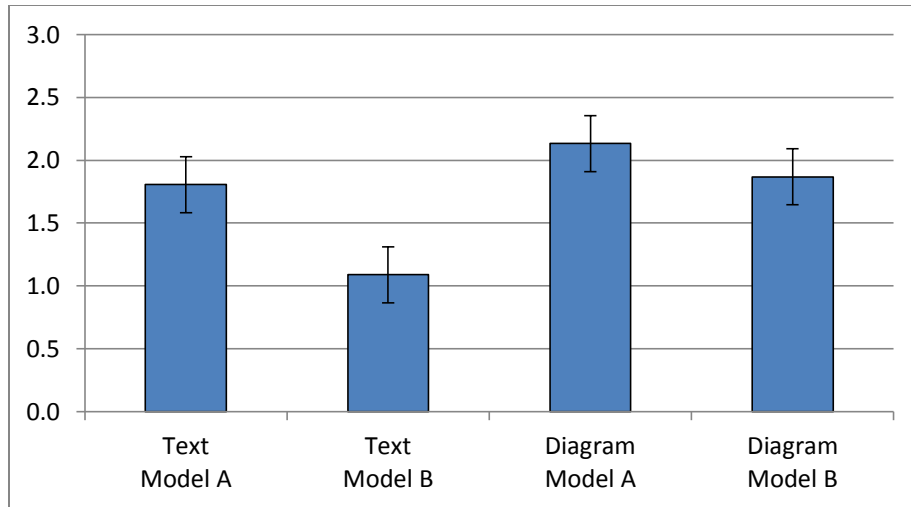


Figure 13. Total Score by condition

Table 17

Mean proportion-correct score for Query by condition

Visualization	Model	Path	Cause	Effect	Expl.
Text	Model A	.77	.91	.87	.32
	Model B	.69	.83	.75	.56
	Marginal Mean	.73	.87	.82	.42
Diagram	Model A	.79	.89	.90	.32
	Model B	.77	.89	.89	.60
	Marginal Mean	.78	.89	.89	.45
Combined	Model A	.78	.90	.89	.32
	Model B	.74	.82	.86	.58
	Mean	.76	.86	.87	.44

Inferences About Causes vs. Inferences About Effects

The primary goal of this study was to examine whether reasoning from cause to effect might result in different performance than reasoning from effect to cause. Model A and Model B were designed so that participants in the Model A conditions were asked about the cause of a variable and participants in the Model B conditions were asked about the effects of the same variable with the answer being identical for both. Ostensibly any differences in performance could be attributed to the causal direction of the question. As Table 17 shows, the relevant comparisons were the Cause Query from Model A with the Effect Query from Model B (i.e. the effects of variable M and the causes of variable M) and the Cause Query from Model B with the Effect Query of Model A (i.e. the effects of variable R and the causes of variable R).

In order to get a single overall test of asymmetries between reasoning from cause to effect and from effect to cause (using both nodes M and R), the data from the Path, Cause, and Effect queries was analyzed using a 2x2 ANOVA with a within-subjects design. The within-subjects factor was Direction of Inference (Cause to Effect, Effect to Cause) and the between-subjects factors were Visualization (Text, Diagram) and Model (Model A, Model B). The Path Query proportion-correct score was used as a covariate to control for possible group differences and model difficulty. Participants scored significantly higher on the Cause Query—i.e., when reasoning from effect to cause ($M = .88, SD = .25$), compared to reasoning from cause to effect in the Effect Query ($M = .86, SD = .27$), $F(1,210) = 7.832, p = .006, \eta_p^2 = .04$. That is to say, participants did better reasoning diagnostically than predictively. There was no significant effect for Visualization, $F(1,210) = 1.952, p = .16$, or for Model, $F(1,210) = 2.357, p = .13$.

Because of differences between text and diagrammed models in terms of ease and accuracy of inferences we also applied this 2x2 within-subjects ANOVA model separately for the Text and Diagram conditions. Table 18 shows the relevant cell means. In the Text conditions, participants scored significantly higher in Cause Query ($M = .87$, $SD = .25$) than in Effect Query ($M = .82$, $SD = .27$), $F(1, 99) = 8.034$, $p = .006$, $\eta_p^2 = .08$. There was also a significant effect for Model; participants with Model A ($M = .89$, $SD = .22$) scored higher than those with Model B ($M = .79$, $SD = .26$), $F(1, 99) = 4.694$, $p = .033$, $\eta_p^2 = .05$. There were no significant differences in the Diagram conditions between Cause Query ($M = .89$, $S.D. = .29$) and Effect Query ($M = .89$, $S.D. = .26$), $F(1, 110) = 1.573$, $p = .21$ or between Model A ($M = .89$, $S.D. = .31$) and Model B ($M = .89$, $S.D. = .23$), $F(1, 110) = .007$, $p = .933$.

Table 18

Mean proportional correct scores for Cause and Effect Queries by condition

Model	Variable M		Variable R		Combined variables		Mean
	A	B	B	A			
Question	Causes	Effects	Causes	Effects	Causes	Effects	
Text	.91	.75	.83	.87	.87	.82	.85
Diagram	.89	.89	.89	.90	.89	.89	.89
Mean	.90	.82	.86	.88	.88	.86	

Discovery of Spurious Correlation

For the Explanation Query, 58% of participants viewing Model B (spurious correlation only) identified the spurious correlation compared to 32% of participants viewing Model A (spurious correlation and direct effect), Wald X^2 (d.f. = 1) = 9.089, $p = .003$. There was no

significant difference between Text (42%) and Diagram (45%) conditions, Wald X^2 (d.f. = 1) = .630, $p = .82$.

Discussion

The main purpose of this study was to examine possible asymmetries in reasoning from cause to effect (predictive reasoning) and from effect to cause (diagnostic reasoning). To examine this, the experiment utilized a causal model and its inverse (all causal directions reversed, still read from left to right). By asking about all the causes of a variable in one condition and all the effects of the same variable in another condition (the answers were identical), this study explored the possibility of a predictive or diagnostic bias.

This experiment found a significant effect for the interaction; however, it was in the opposite direction of the hypothesis. Participants did better at diagnostic reasoning. In other words, participants did better reasoning about the cause from the effect even though it is in the opposite direction of temporal flow and, additionally, required reading the model from right to left. It is important to note that the differences in performance between reasoning from cause to effect(s) and from effect to cause(s) were entirely within the Text conditions. The Diagram conditions scored almost identically. Because this finding was unexpected, prior to further discussion, another experiment was run in an attempt to replicate these results. Further discussion about predictive and diagnostic reasoning can be found in Study 4.

Diagram vs. text. This study replicated findings from the previous studies that show participants answer questions more accurately and in less time when presented with a diagrammed version of a causal model. Participants in the diagram conditions were 40% faster and scored 25% higher. Additionally, the diagram appeared to negate any bias to reasoning that

occurred due to predictive or diagnostic reasoning. There was no difference in the performance of participants in the diagram condition.

Differences between models. The models were created to be functionally equivalent and appeared to be of equal complexity. Because the models were constructed of abstract variables, there was no reason to believe one model would be more difficult than the other. However, there was a significant difference in performance due to the model participants used. Participants in the Text/Model B condition scored the lowest on each of Path, Cause, and Effect Queries. Because this difference was consistent across the different types of tasks, it would seem that the effect was due to the text version of Model B having a higher inherent difficulty. It is unclear why this effect occurred and why it only manifest in the Text condition.

One possible explanation for this difference may be found in the nature of the relationships in the model. Taken as a whole, the relationships were fundamentally identical between the two models. The causal chains were identical as well. However, individual variables differed in the complexity of chains leading in the two possible directions. In Model A, the number of effects each variable respectively had was 3, 3, 2, 1, and 1. In Model B, it was: 4, 2, 2, 1, and 1. Although the number of relationships in the model was the same (10 dichotomous relationships), the proportion of effects was slightly different, specifically, the first two variables in the model. In Model A, the first two variables have three effects each. In Model B, the first variable has four effects and the second has 2 effects. It may not seem like a substantial difference but when variables are presented as text, small changes can greatly increase complexity. The diagram of the same model may have provided enough of an advantage to prevent from affecting participants.

Finding alternative explanations. The Explanation Query asked participants to explain the relationship between two variables, with the correct answer being that it is a spurious correlation due to a common cause. In previous studies, participants were asked to explain the relationship between two variables, but the answer was much more complex and only a small number of participants identified the spurious correlation (roughly 16%). In Model A of this study, the variables had a direct relationship (i.e., one variable affected the other directly) but both variables also shared a common cause. In Model B, there was no direct relationship between the variables, but they again shared a common cause.

It appears the simplification was successful as substantially more participants discovered the spurious correlation. Participants in Model B conditions (the conditions without the direct effect) identified the spurious correlation 58% of the time compared to 32% of the time in Model A conditions. It appears that this difference could be explained similarly to the blocking-type error discussed in the introduction, i.e., participants discovered the direct affect and may have ceased searching for further explanation. Nevertheless, it is encouraging that people with little statistical background were so often able to discover the concept of a spurious correlation without instruction. Diagrams did not appear to help, since participants in both Text and Diagram conditions found the spurious correlation at a similar rate. There was an effect on work time, however, as participants with diagrams completed the task about 30 seconds faster.

Participants often discovered the spurious correlation while simultaneously demonstrating other misconceptions about statistics and causality in general. This is to be expected with a sample that overall did not have much formal education in statistics. However, one error was produced with surprising frequency. The most common misconception, demonstrated by a large portion of participants who identified a spurious correlation, was that a

positive correlation between variables could be explained by their common cause, but also by their common effect. That is to say, since both variables affect a third variable, that would explain the first two variables' positive correlation. This “intuitive” understanding of causal relationships may provide insight into how people naively reason using causal models.

Study 5: Replication of Asymmetry Study

The results of the previous experiment indicated a bias towards diagnostic reasoning over predictive reasoning. This seemed to contradict previous findings that predictive reasoning was an easier process (Tversky & Kahneman, 1980; Waldmann & Holyoak, 1992). Study 5 was designed as a replication of Study 4 in order to further investigate possible asymmetries in reasoning. A new, more complex causal model was introduced.

Another purpose for replicating Study 4 was to address possible bias the phrasing of the questions may have introduced. In Study 4, the Cause Query was phrased in the active voice (“What variables *affect*...””) and the Effect Query was phrased in the passive voice (“What variables *are affected by*...”). This new study attempts to rectify that possible source of bias by phrasing both Queries in the active voice. The Effect Query was changed to “What variables does X *affect*?”

Methods

Participants

We recruited 240 participants from Amazon’s Mechanical Turk (MT) for this study. The experiment was only available to participants who had an American IP address. Two participants were excluded for not completing the task. Another 23 participants were excluded for having participated in one of the three previous studies. This resulted in 215 participants randomly assigned to four conditions.

The participants were 60% male. English was the native language for 96% of participants. The mean age was 32 years old ($SD = 11$) and ranged from 18 to 84 years. About 82% of participants reported having attended at least some college with 18% having attended at least some graduate school. Of those who attended college, 23% (19% overall) majored in

mathematics, engineering, computer science, or information systems. However, only 3% of all participants reported ever having taken more than two courses in statistics. About half (49%) reported never having taken a statistics course.

Stimuli

Participants were presented with an abstract causal model that consisted of two more nodes than the models in the previous studies (albeit the number of dichotomous causal relationships remained the same). Participants either saw this model (Model A) or its inverse (Model B). The inverse model had all causal link directions reversed, and was reflected around the vertical axis so that the general direction of causal flow remained from left to right. The models were also presented as either Text or as a Diagram. Please see Figure 14 for the stimuli used in this study.

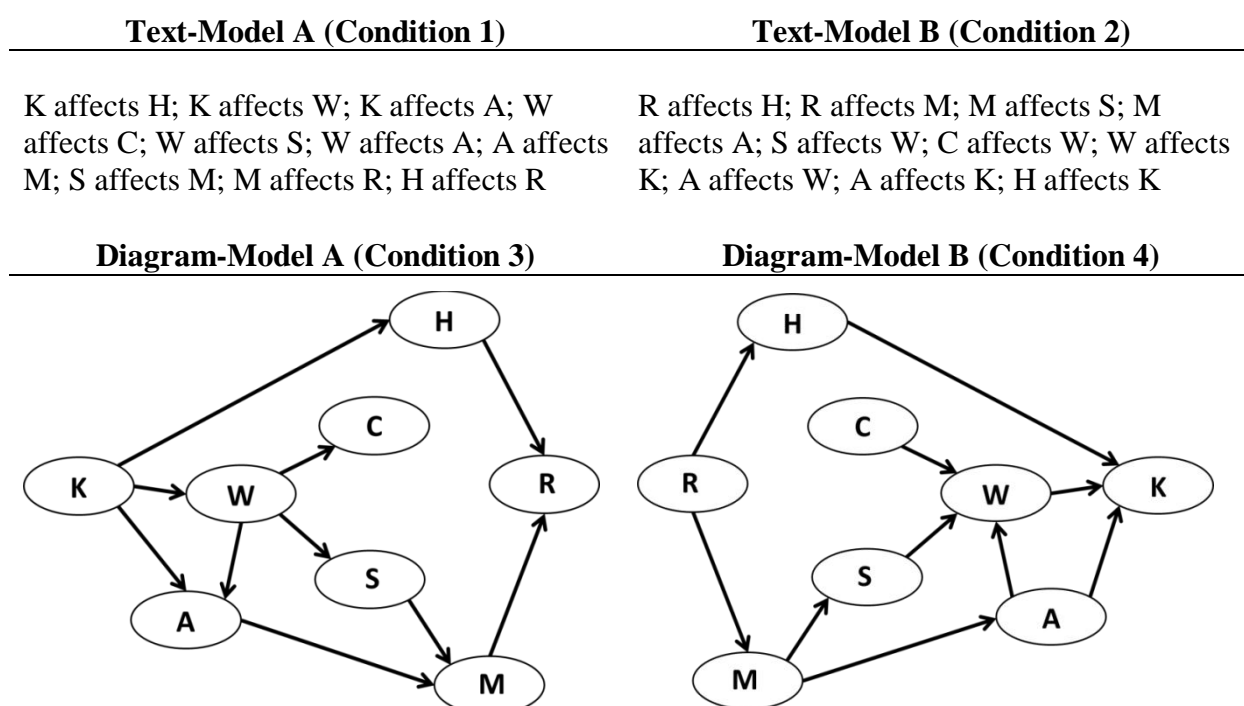


Figure 14. The four conditions for displaying the causal model in Study 5.

Procedure

Our task was hosted online by MT. It was written in HTML and JavaScript. Participants were paid \$1.00 to complete the task. Once participants accepted the task, they were randomly assigned to one of the four conditions and presented with a short introduction. They were allowed a maximum of one hour from the time they accepted the task to fully complete and submit it. Following is the introduction for all conditions:

A primary goal of science is to uncover the causes of the phenomena of interest. Because many phenomena have multiple causes, some direct and some indirect, uncovering causal relationships can be complicated. For example, food consumption is a direct cause of obesity. The number of hours of TV watched per week is an indirect cause because greater TV time reduces energy expenditure. After preliminary research, scientists often model the causal relations they have found, and use the model to come to conclusions.

Consider a simple example, represented by the information below:

X affects Y, X affects Z, and Y affects Z.

*In this case, X has a causal influence on Z in two ways. First there is a **direct effect** of X on Z (by assumption). Also there is an **indirect effect** of X on Z, because X is assumed to affect Y and Y is assumed to affect Z. To summarize, if you were given this information*

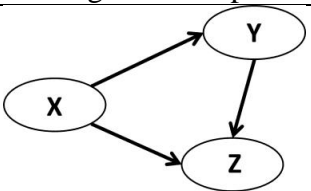
Text example	Diagram example
<p><i>X affects Y, X affects Z, and Y affects Z.</i></p>	 <pre> graph LR X((X)) --> Y((Y)) X((X)) --> Z((Z)) Y((Y)) --> Z((Z)) </pre>

Figure 15. The example models for Study 5

The participant was then presented with the example causal model. Depending on the condition to which they had been assigned, it was either in text or diagram form (see Figure 15).

The instructions continued:

and asked to list all the ways that variable X causally affects variable Z, you would write:

X affects Z.

X affects Y which affects Z.

After the introductory example, participants clicked the “proceed” button and saw the following task description and the target causal model:

Assume that a researcher makes causal assumptions about a particular social science domain where variables C, R, H, M, A, and S are measurable aspects of people.

The researcher makes the following causal assumptions:

The causal model, in one of the four versions shown in Figure 14, was presented at this point. With the preceding text and the model visible on the screen, participants were presented with the first of four questions. After they answered that question, they pressed a button and the first question (along with their answer) disappeared and the second question appeared. Meanwhile, the model remained on the screen. This was repeated until the final question which, upon completion, directed participants to a new page asking for demographic information. From this page they submitted their entire task. Participants were not permitted to go back to a previous screen to view or change answers.

Outcome measures

Answers were collected from the four questions presented to the participants. This study differed from the previous studies in that the position of the Path Query was moved from the first question to the last. Also, the Alternative Explanation Query was not included in this study. The

first and second questions (Cause Query and Effect Query respectively) asked participants to identify causes and effects of certain variables to examine possible asymmetries in reasoning from cause to effect and from effect to cause. To control for practice and other order effects, the order in which these two questions appeared was randomized for each participant. The third question (Path Query) asked participants to identify all direct and indirect effects between two variables. Table 19 has the specific wording of the questions and their respective answers.

Table 19

Study 5 tasks and answers

Query	Model A		Model B	
	Question Text	Answer	Question Text	Answer
Cause	Please list <u>all</u> the VARIABLES that AFFECT variable M (just name the variables that are <u>causes</u> of M, don't list paths).	K, W, A, S	Please list <u>all</u> the VARIABLES that AFFECT variable W (just name the variables that are <u>causes</u> of W, don't list paths).	A, C, M, S, R
Effect	Please list <u>all</u> the VARIABLES that variable W AFFECTS (just name the variables that are <u>effects</u> of W, don't list paths).	A, C, M, S, R	Please list <u>all</u> the VARIABLES that variable M AFFECTS (just name the variables that are <u>effects</u> of M, don't list paths).	K, W, A, S
Path	Please list <u>all</u> the ways that variable K could affect variable M.	K->A->M K->W->A->M K->W->S->M	Please list <u>all</u> the ways that variable M could affect variable K.	M->A->K M->A->W->K M->S->W->K

We analyzed the data in two ways. If participants listed all of the correct paths/variables and did not list any incorrect paths/variables, the question was scored as correct. A participant's Total Score was the sum of all their correct answers.

However, because each question was composed of multiple answers, it was possible to answer several parts of a question correctly yet still not get the entire question correct.

Therefore, to allow partial credit, a second dependent measure was created. This variable was created by summing the number of correct paths or variables each participant listed for their response, subtracting the number of incorrect answers, and then dividing that sum by the number of possible correct answers. This created a variable for each of the first three questions that we called Proportion-correct Score (PropScore). The maximum score for this variable is 1 (indicating that the answer is entirely correct); the minimum score could be negative.

Results

Participants completed the entire task with a mean response time of 9.7 minutes and a median response time of 4.4 minutes. Because the presence of outliers strongly affected the mean, median time may be more interpretable. A 2x2 ANOVA was run for Total Time using Visualization and Content. There was not a significant difference between Text and Diagrams, $F(1, 157) = 0.41, p = .521$ or Model A and Model B, $F(1, 157) = 0.54, p = .465$. A breakdown of median response times for each question by condition can be found in Table 20.

Two research questions guided our analyses of accuracy of inferences. The first question was whether the presentation form of the causal model, either text or diagram, would affect successful interpretation of the model. The second is whether participants would show asymmetry in how successfully they reasoned from cause to effect compared to from effect to cause.

Table 20

Median response time (in seconds) by condition for Study 5 Queries

Visualization	Model	Cause	Effect	Path	Total Time	N
Text	Model A	80	55	94	340	57
	Model B	78	43	101	319	48
	Text Median	80	51	99	324	105
Diagram	Model A	69*	28*	52	223	56
	Model B	56*	24*	47	216	54
	Diagram Median	61	27	50	221	110
Combined	Model A	76	42	72	281	113
	Model B	70	39	66	265	102
	Median	74	40	70	265	215

Text vs. Diagram

In order to assess the overall effects on inference performance of using a causal diagram versus text, we analyzed the Total Score variable using a 2x2 ANOVA with the following factors: Visualization (Text, Diagram); Model (Model A, Model B). The participants in the Diagram conditions scored significantly higher ($M = 1.8$, $SD = 1.0$) than in the Text conditions ($M = 1.1$, $SD = 1.1$), $F(1,211) = 20.659$, $p < .001$, $\eta_p^2 = .10$. There was not a significant score difference between Model A ($M = 1.5$, $SD = 1.1$) and Model B ($M = 1.4$, $SD = 1.2$), $F(1,211) = 1.906$, $p = .17$. The Visualization*Model interaction was not significant, $F(1,211) = 0.004$, $p = .95$. Only six participants (five from the Text conditions) drew their own external aids (e.g. diagrams or tables). Please see Table 21 and Figure 16 for further details.

* A programming bug prevented time data from being collected for the first two questions of participants in the Diagram conditions that saw the Effect Query first. Therefore, these numbers do not accurately reflect the size of the time difference. However, the pattern of the Cause Query taking longer than the Effect Query was exhibited in the Text conditions regardless of which Query came first so it may be safe to assume the same for the Diagram condition.

Table 21

Descriptive statistics for all four conditions

Visualization	Model	Total Score Mean	S.D.	N
Text	Model A	1.21	1.0	57
	Model B	1.02	1.1	48
	Marginal Mean	1.12	1.1	105
Diagram	Model A	1.88	1.0	56
	Model B	1.67	1.1	54
	Marginal Mean	1.77	1.0	110
Combined	Model A	1.54	1.1	113
	Model B	1.36	1.2	102
	Mean	1.46	1.1	215

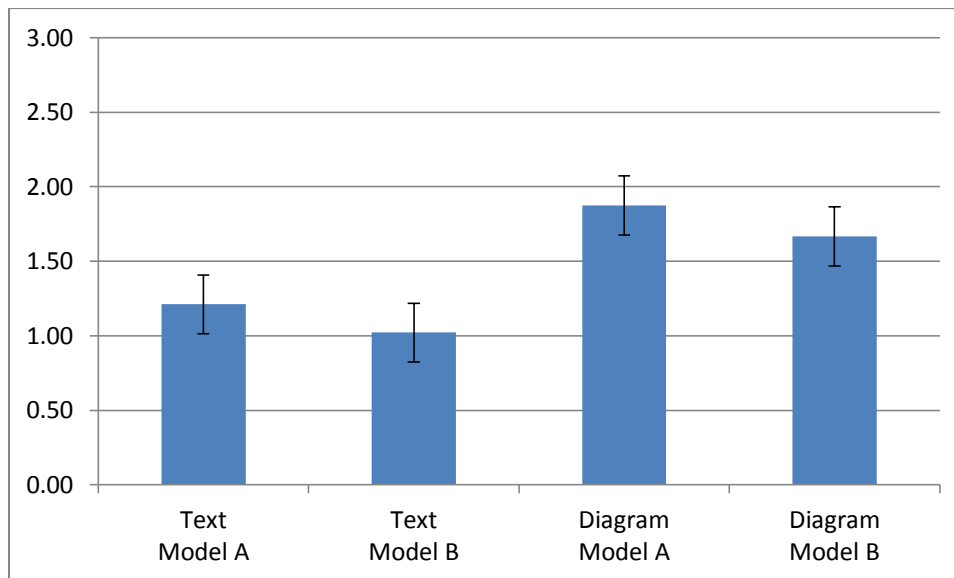


Figure 16. Total Score by condition

Table 22

Mean proportion-correct score for Query by condition

Visualization	Model	Cause	Effect	Path
Text	Model A	0.75	0.75	0.49
	Model B	0.67	0.68	0.53
	Marginal Mean	0.71	0.72	0.51
Diagram	Model A	0.76	0.78	0.74
	Model B	0.79	0.76	0.70
	Marginal Mean	0.77	0.77	0.72
Combined	Model A	0.75	0.77	0.61
	Model B	0.73	0.72	0.62
	Mean	0.74	0.75	0.62

Inferences About Causes vs. Inferences About Effects

The primary goal of this study was to replicate the findings from Study 4 that reasoning from cause to effect (predictive reasoning) resulted in different performance than reasoning from effect to cause (diagnostic reasoning). As Table 22 shows, the relevant comparisons were the Cause Query from Model A with the Effect Query from Model B (i.e. the effects of variable M and the causes of variable M) and the Cause Query from Model B with the Effect Query of Model A (i.e. the effects of variable W and the causes of variable W).

In order to get a single overall test of asymmetries between reasoning from cause to effect and from effect to cause (using both nodes M and W), the data from the Cause and Effect queries were analyzed using a 2x2 within-subjects ANOVA. The within-subjects factor was Direction of Inference (Cause to Effect, Effect to Cause) and the between-subjects factors were

Visualization (Text, Diagram) and Model (Model A, Model B). Participants did not score significantly higher on the Cause Query—i.e., when reasoning from effect to causes ($M = .74$, $SD = .36$), compared to reasoning from causes to effects in the Effect Query ($M = .75$, $SD = .35$), $F(1,211) = 0.014$, $p = .91$. That is to say, participants did no better reasoning predictively than diagnostically. There was not a significant effect for Visualization, Text ($M = .72$, $SD = .33$), Diagram ($M = .77$, $SD = .37$), $F(1,211) = 2.142$, $p = .15$ or for Model, Model A ($M = .76$, $SD = .35$), Model B ($M = .71$, $SD = .34$), $F(1,211) = 0.904$, $p = .34$.

Because of differences between text and diagrammed models in terms of ease and accuracy of inferences, the Text and Diagram conditions were analyzed separately using a 2x2 within-subjects ANOVA. Table 23 shows the relevant cell means. In the Text conditions, participants did not score significantly higher in Cause Query ($M = .71$, $SD = .34$) than in Effect Query ($M = .72$, $SD = .33$), $F(1,103) = 0.005$, $p = .064$, $\eta_p^2 = .80$. There was no significant difference between participants with Model A ($M = .75$, $SD = .32$) and those with Model B ($M = .67$, $SD = .35$), $F(1,103) = 2.209$, $p = .14$.

There was no significant difference in the Diagram condition between Cause Query ($M = .77$, $S.D. = .37$) and Effect Query ($M = .77$, $S.D. = .38$), $F(1,108) = 0.005$, $p = .94$. Neither was there a significant difference between Model A ($M = .77$, $S.D. = .39$) and Model B ($M = .78$, $S.D. = .32$), $F(1,108) = 0.004$, $p = .95$.

Table 23

Mean proportional correct scores for Cause and Effect Queries by condition

Model	Variable M		Variable W		Combined variables		Mean
	A	B	B	A			
Question	Causes	Effects	Causes	Effects	Causes	Effects	
Text	0.75	0.68	0.67	0.75	0.71	0.72	0.72
Diagram	0.76	0.76	0.76	0.78	0.77	0.77	0.77
Mean	0.75	0.72	0.73	0.76	0.74	0.75	

Discussion

Study 5 was designed as a replication of Study 4 to examine possible asymmetries in reasoning between causes and effects. Study 4 had shown a tendency for people to do better on an inference task when asked to reason about causes from effects as opposed to reasoning about effects from causes. This finding was only demonstrated when the causal model was in text form. With a causal model in diagram form, participants did equally well in either direction.

Study 5 did find a similar effect, but only for one of the causal models (Model A). When the causal flow in the model was reversed (creating Model B), the effect reversed itself as well, i.e., participants did better reasoning from cause to effect. The difference in scores was such that they effectively cancelled each other out so that there was no main effect of Direction of Inference.

As before, these differences were confined to the Text conditions. Consistent with the results of our previous studies, participants presented with models using diagrams outperformed participants with text. However, it is interesting to note that the scores for the questions in which participants did better in the Text conditions (i.e. the cause Query for Model A and the Effect Query for Model B) were roughly equivalent to the scores in the Diagram conditions. One way

to interpret this is to conclude that the advantage diagrams offered was not to allow participants to outperform the Text conditions, but rather to prevent errors or difficulties prevalent in the Text conditions.

The inability to replicate the previous study's findings regarding asymmetrical facility in reasoning about causes and effects when using a formal causal model suggests that findings in which people demonstrate differences when reasoning about causes from effects and vice versa may be due more to the specific structure of the model than a cognitive bias. A consistent finding across the two studies was that regardless of the model and the direction of inference, participants in the diagram conditions scored almost identically. Any asymmetries between predictive and diagnostic reasoning that might exist may be mitigated by the presence of an externally represented causal model.

General Discussion

Many scientific fields make use of formal causal modeling techniques to reason about and then communicate findings. Despite the proliferation of diagrammatic representation of causal models, little research has been conducted on whether diagrams even help facilitate understanding, and if they do, what the essential features of diagrams are. The findings from studying effective causal model representation may be of use to instructors teaching techniques for creating formal models to students and may also be of use to recipients of formal models who themselves have little training in formal modeling techniques.

This set of experiments examined four aspects of the use of diagrams for presenting causal models. Each experiment investigated performance differences in deductive inference between causal models presented as diagrams and presented as text. The first experiment examined the impact of specific content on reasoning about causal models. The second and third experiments examined the benefits of specific aspects of diagrams—spatial layout and the use of arrows—for diagrammed causal models, in particular the extent to which they facilitate successful reasoning. The fourth and fifth experiments examined the possibility of differences in inference performance between predictive and diagnostic reasoning, and whether any such differences might be affected by diagram use

Text vs. diagrams

Research has shown the benefit of diagrams over text for post-test knowledge and inference questions (Ainsworth & Loizou, 2003), training for a future task (Kaminsky, Sloutsky, & Heckler, 2008), and strategy selection (Petre & Green, 1993). This advantage for diagrams over text is hardly one that holds across all situations. The benefits of diagrams are often found to be require expertise in a particular field (Hegarty & Sims, 1994; Heiser & Tversky, 2002;

Kaminsky, Sloutsky, & Heckler, 2008; Petre & Green, 1993; Suwa & Tversky, 1997) especially in functional (vs. structural) diagrams. Everyone has experience with informal causal models, so in a sense, everyone is an expert. However, constructing a formal causal model is not a familiar experience because it requires expertise in statistics, so in a sense, many people are actually novices. Would lack of expertise with formal causal models negate any diagrammatic benefit that causal models may provide?

The major finding from this set of experiments is that presenting a causal model as a diagram resulted in better outcomes much more often than did using a text version of the same model. The current results found advantages for diagrams in the speed of task completion, the number of results, the thoroughness of responses, and the minimization of errors found in text versions.

Participants with diagrams completed every task with a faster median work time than participants with text models, although the difference was not always significant. In some cases, a Query in the Diagram condition was completed twice as fast as in the Text condition. The task where this finding was most consistent was in the Path Query where the time difference was significant in each study. This is also the Query that took the most time to complete in each experiment.

Additionally, this was not a case of a speed/accuracy tradeoff as these faster times did not result in lower scores. Participants with diagrams outperformed participants with text in almost every task. The one exception to this was when arrowheads were replaced with words; the effect of which will be discussed later in the section on model structure. The effect sizes were small and the difference was not always significant, but the finding was consistent across every other Query.

The manner in which participants scored higher is worth noting. Errors were rarely made regardless of the experiment or condition within an experiment. Rather the advantage for the diagram condition came in the form of participants discovering more answers. Each task had several components to it and the diagram conditions seemed to facilitate finding more answers and in less time.

Diagrams also seem to mediate complexity. In the experiments on predictive and diagnostic reasoning, the type of reasoning was not as influential on performance as was the specific composition of the model. This effect was only present in the Text conditions however. Any obstacle to reasoning that arose in the Text conditions was not present in the Diagram conditions.

There was one area where diagrams did not aid reasoning and two areas where they actually impeded reasoning. The area where diagrams did not aid reasoning was in the discovery of alternative explanations for variable relationships. In the first two studies, participants were given two variables and asked to explain how they might be related. The answer involved discovering five possibilities ranging from direct and indirect effects to a common cause. This proved to be a difficult task, with only a small percentage of participants discovering the common cause. In the third study, the task was simplified so that in one condition there were only two possibilities: a common cause and a direct effect. About one third of participants found the common cause in this condition. The other condition had only a common cause; here, almost two-thirds of participants found the common cause. This difference between conditions was expected because the models were designed to create a blocking effect where the salience of the direct effect discouraged further exploration. Blocking in this instance is defined as a stimulus variable paired with a response and then a second stimulus variable paired with the same

response resulting in the second variable not being learned as an alternative stimulus, or in this case, cause (Kamin, 1969). The surprising result was that this error occurred in the Diagram condition as frequently as in the Text condition. The spatial organization in the diagram was such that the arcs indicating the direct affect and the common cause were laid out right next to each other and required minimal search (compared to the Text condition) to discover both. It seems that in this situation perhaps the blocking effect overrode the benefits of a diagram for a statistically inexperienced sample.

Blocking is not the only explanation, however. Mackie's (1965) assertion that causal reasoning is frequently identification of INUS variables (variables that are necessary for an effect but only within a condition that itself is unnecessary) may bear on these findings. An INUS variable, by definition, is an attempt to isolate a singular cause amongst a field of other causes, all of which are sufficient to bring about the effect. That is to say, it is the process of assigning causal attribution to one event rather than parsing out different weights to the entirety of the causal field, as is done in path analysis. This method of isolating causes with informal causal reasoning may conflict with the method of formal causal reasoning which is to identify the entire range of causes. Both blocking and INUS reasoning would produce the same pattern of answers where difficult, complex, or redundant answers are overlooked.

Content of the causal model

One of the main exceptions to the advantage of diagrams over text was found in the first study that examined the effect of causal model content. While an abstract diagram outperformed the text conditions at approximately the same rate as in other studies, the diagram comprised of real-world content scored equivalently to the Text condition. Task time was equivalent between all conditions with the exception that participants given abstract diagrams performed the path

finding task faster. This is likely due in large part to the greater amount of writing required for the answer. That is to say, that writing the answer in the Abstract conditions consisted of writing “A”, “B”, “C”, etc. and writing the answer in the Concrete conditions consisted of writing “combat duration”, “received injury”, “psychological stress”, etc. All other things being equal, writing more will take more time. It does not appear that the concrete context aids reasoning, and may in fact, hinder the sort of deductive inference task (involving mainly paths search) examined here.

The inferior performance for participants with concrete diagrams seemed to stem from errors of omission rather than errors of commission. Incorrect answers were equally rare between all conditions, in contrast to other research that indicates performance deficits are due to real-world variables (Cummins et al., 1988; Geary, 1994).

The fundamental difference between formal and informal, specifically INUS reasoning, may explain these results. As previously explained, the search for causes can be categorized into identifying one cause or multiple causes. One of the logical errors discussed in Chapter Two was ignoring multiple causes. INUS reasoning seeks to identify the necessary event within a field of sufficient but unnecessary other causes (the causal field). Although INUS reasoning does not seek to ignore multiple causes, it seems feasible, given people’s reputation as being “cognitive misers” (Fiske & Taylor, 1984), that identifying the INUS variable could transform from identifying the *necessary* cause to simply identifying the cause.

Formal methods for causal reasoning differ in that they seek to identify multiple causes and to assign weight to each cause based on how much each causal variable contributes to the overall effect. Thinking about what brought about an effect in this way, identifying the causal

field, is fundamentally different than thinking in an informal INUS manner, attempting to identify a single necessary cause.

Under the assumptions of dual-process theories, this difference between types of reasoning may not normally matter because informal causal reasoning would be handled by System I and formal causal reasoning would fall under System II. There may be some features of the causal models used in these experiments however, that place them on the cusp between implicit and explicit reasoning. Research has shown that when diagrams are analytically studied, they can facilitate higher-order thinking (Cromley et al., 2010) but often diagrams are approached and interpreted using System I (Alter et al., 2007; Cromley et al. 2010). The abstract content of a diagrammed causal model may be difficult enough in appearance to cue System II (Alter et al., 2007) whereas the familiar content in a concrete causal model may not do the same, leaving the analysis to System I processing, and subsequently, a less thorough analysis of the model.

If this is indeed the process that resulted in abstract diagrams enabling better performance, it may be beneficial to introduce some “desirable difficulty” (Bjork, 1994) to cue System II. The other benefits of diagrams, such as spatial layout, may improve accuracy enough to make up for the efficiency lost by not using the intuitive, implicit System I.

Structure of the causal model

Two studies replacing arrowheads with words and vice versa indicated that this manipulation did not seem to affect reasoning about causal models when they were presented as text. But it did seem to affect reasoning using causal diagrams. Even by varying the position of the words and testing participants to ensure understanding prior to the task, participants given

words instead of arrows in a causal diagram often interpreted the words to mean bidirectional causality.

There are several reasons to doubt that participants truly believe that the links are bidirectional. Words are as asymmetrical as arrows. The words were placed in the same position as arrowheads. And participants had already passed a test acknowledging that the words did not indicate bidirectional relationships. Rather these findings seem to offer evidence that diagrams activate intuitive reasoning. This activation may be one of the reasons diagrams are so useful and efficient in conveying information.

As previously discussed, diagrams consistently outperform text in both speed and accuracy. Diagrams perhaps present information in a way that allows for subconscious processes to reason about the information contained therein. However, this same advantage can be a detriment if some feature of the diagram triggers an unconscious response that leads to systematic bias, much as was observed within these experiments where participants frequently made an irrational error.

Predictive vs. diagnostic reasoning

These studies did not produce any evidence for differences in the ease or accuracy of predictive or diagnostic reasoning. No such bias appeared when participants were presented with a diagrammed causal model. In the Text conditions, no consistent bias appeared. The structure of the model seemed to be more of a factor in determining whether predictive or diagnostic reasoning was more successful. Given that previous work in this area was always done using grounded, contextual problems, it is reasonable to conclude that any advantage of predictive over diagnostic reasoning is due to the content of the problem rather than a cognitive preference or differential ease for one type of reasoning.

Conclusion

The findings across these studies suggest that causal models presented as diagrams have certain properties that aid in reasoning about causality. Diagrams are useful because they can use space to present information in a manner that facilitates organization and inference. In a variety of tasks, participants with diagrams consistently outperformed counterparts presented with text. The exceptions to this overall effect may provide insight into processes associated with using diagrams to reason about causality.

The pattern of results also seems to indicate that diagrams may activate intuitive and heuristic System I reasoning. Participants presented with Concrete content were not as thorough when answering questions and participants presented with words in the place of arrows did much poorer than those with traditional arrows. Although often advantageous, this type of reasoning is susceptible to errors that would not appear if people were thinking more analytically.

Diagrams did not prevent a blocking effect from occurring when participants were tasked with providing alternative explanations for variable relationships. Additionally, diagrams that violate traditional construction, in this case the use of words to show directionality instead of arrowheads, can dramatically interfere with performance. Finally, the use of contextually grounded variables in a diagrammed causal model may be a hindrance when compared to an equivalent abstract model. The presence of these errors indicates further study is needed in order to more fully understand what cognitive processes are occurring when people reason with causal models.

References

- Ahn, W. (1998). Why are different features central for natural kinds and artifacts. *Cognition*, 69, 135-178.
- Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, 54, 299-352.
- Ahn, W., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology*, 41, 361-416.
- Ainsworth, S., & Loizou, A. T. (2003). The effects of self-explaining when learning with text or diagrams. *Cognitive Science*, 27(4), 669-681.
- Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General*, 136(4), 569-576.
- Aristotle. (n.d./2004). *Metaphysics*. (W. D. Ross, Trans.) London: Penguin Books Ltd.
- Bacon, J., Campbell, K., & Reinhardt, L. (Eds.). (1993). *Ontology, Causality, and Mind*. Cambridge: Cambridge University Press.
- Baranes, R., Perry, M., & Stigler, J. W. (1989). Activation of real-world knowledge in the solution of word problems. *Cognition and Instruction*, 6(4), 287-318.
- Bauer, M. I., & Johnson-Laird, P. N. (1993). How diagrams can improve reasoning. *Psychological Science*, 4(6), 372-378.
- Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe, & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185-205). Cambridge, MA: MIT Press.
- Bohm, D. (1957). *Causality and Chance in Modern Physics*. Philadelphia: University of Pennsylvania Press.
- Bollen, K. A., & Pearl, J. (2012). Eight myths about causality and structural equation modeling. In S. Morgan (Ed.), *Handbook of Causal Analysis for Social Research*. Springer.
- Buehner, M. J., & Cheng, P. W. (2005). Causal learning. In K. J. Holyoak, & B. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 143-168). Cambridge, England: Cambridge University Press.
- Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, 104(2), 367-405.

- Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, 17(4), 391-416.
- Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology*, 58(4), 545-567. doi:10.1037/0022-3514.58.4.545
- Cheng, P. W., Novick, L. R., Liljeholm, L. R., & Ford, C. (2007). Explaining four psychological asymmetries in causal reasoning: Implications of causal assumptions for coherence. In J. K. Campbell, M. O'Rourke, & H. Silverstein (Eds.), *Causation and Explanation* (pp. 1-32). Cambridge, MA: MIT Press.
- Cohen, L. B., Amsel, G., Redford, M. A., & Casasola, M. (1998). The development of infant causal perception. In L. B. Cohen, G. Amsel, M. A. Redford, M. Casasola, & A. Slater (Ed.), *Perceptual Development: Visual, Auditory, and Speech Perception in Infancy* (pp. 167-209). East Sussex, UK: Psychology Press.
- Corter, J. E., Mason, D. L., Tversky, B., & Nickerson, J. V. (2011). Identifying causal pathways with and without diagrams. In L. Carlson, C. Hoelscher, & T. Shipley (Ed.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2715-2720). Austin, TX: Cognitive Science Society.
- Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, 31(3), 187-276.
- Cromley, J. G., Snyder-Hogan, L. E., & Luci-Dubas, U. A. (2010). Cognitive activities in complex science text and diagrams. *Contemporary Educational Psychology*, 35(1), 59-74.
- Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology*, 20(4), 405-438.
- Day, S. B., Manlove, S., & Goldstone, R. L. (2011). Transfer, and the effects of context outside of the training task. *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society* (pp. 2637-2642). Boston, Massachusetts: Cognitive Science Society.
- Dean, T., & Kanazawa, K. (1989). A model for reasoning about persistence and causation. *Computational Intelligence*, 5(2), 142-150.
- Easterday, M. W., Aleven, V., & Scheines, R. (2007). 'Tis better to construct than to receive? The effects of diagramming tools on causal reasoning. In R. Luckin, K. Koedinger, & J. Greer (Eds.), *Proceedings of the 13th International Conference on Artificial Intelligence in Education*. Amsterdam: IOS Press.

- Easterday, M. W., Aleven, V., Scheines, R., & Carver, S. M. (2009). Constructing causal diagrams to learn deliberation. *International Journal of Artificial Intelligence in Education*, 19(4), 425-445.
- Ericsson, K. A., Chase, W. G., & Faloon, S. (1980). Acquisition of a memory skill. *Science*, 208(4448), 1181-1182.
- Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(3), 678-693.
- Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General*, 140(2), 168-185.
- Ferreira, F. (1994). Choice of passive voice is affected by verb type and animacy. *Journal of Memory and Language*, 33, 715-736.
- Fiske, S. T., & Taylor, S. E. (1984). *Social Cognition*. Reading, MA: Addison-Wesley Publishing Company.
- Frede, M. (1987). The original notion of cause. In M. Frede, *Essays in Ancient Philosophy* (pp. 125-150). Minneapolis: University of Minnesota Press.
- Freedman, D. A. (2007). Statistical models for causation. In W. Outhwaite, & S. Turner (Eds.), *The SAGE Handbook of Social Science Methodology* (pp. 127-146). Sage Publications Ltd.
- Gauss, C. F. (1809). *Theoria Motus Corporum Coelestium*. Hamburg: Perthes et Besser.
- Geary, D. C. (1994). *Children's Mathematical Development: Research and Practical Applications*. Washington D.C.: American Psychological Association.
- Gobert, J. D., & Clement, J. J. (1999). Effects of student-generated diagrams versus student-generated summaries on conceptual understanding of causal and dynamic knowledge in plate tectonics. *Journal of Research in Science Teaching*, 36(1), 39-53.
- Goswami, U., & Brown, A. B. (1990). Melting chocolate and melting snowmen: Analogical reasoning and causal relations. *Cognition*, 35(1), 69-95.
- Hagmayer, Y., & Waldmann, M. R. (2000). Simulating causal models: The way to structural sensitivity. In L. Gleitman, & A. Joshi (Ed.), *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.
- Hegarty, M. (1992). Mental animation: Inferring motion from static diagrams of mechanical systems. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18, 1084-1102.

- Hegarty, M., & Sims, V. K. (1994). Individual differences in mental animation during mechanical reasoning. *Memory & Cognition*, 22(4), 411-430.
- Heiser, J., & Tversky, B. (2002). Diagrams and descriptions in acquiring complex systems. In W. D. Gray, & C. D. Schunn (Ed.), *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 447-452). Mahwah, NJ: Lawrence Erlbaum Associates.
- Horn, R. E. (1998). *Visual Language: Global Communication for the 21st Century*. Bainbridge Island: MacroVu, Inc.
- Hume, D. (1739/2000). *A Treatise of Human Nature*. (D. F. Norton, & M. J. Norton, Eds.) USA: Oxford University Press.
- James, C. T. (1975). The role of semantic information in lexical decisions. *Journal of Experimental Psychology: Human Perception and Performance*, 104, 130-136.
- Kamin, L. J. (1969). Predictability, surprise, attention and conditioning. In B. A. Campbell, & R. M. Church (Eds.), *Punishment and Adversive Behavior* (pp. 279-296). New York: Appleton-Century-Crofts.
- Kaminsky, J. A., Sloutsky, V. M., & Heckler, A. F. (2008). The advantage of abstract examples in learning math. *Science*, 320, 454-455.
- Kant, I. (1781/1999). *Critique of Pure Reason (The Cambridge Edition of the Works of Immanuel Kant)*. (P. Guyer, & A. W. Wood, Eds.) Cambridge, UK: Cambridge University Press.
- Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, 28(2), 107-128.
- Kieras, D. (1978). Beyond pictures and words: Alternative information-processing models for imagery effects in verbal memory. *Psychological Bulletin*, 85, 532-554.
- Koedinger, K. R., & Nathan, M. J. (2004). Effects of representation on quantitative reasoning. *the Journal of the Learning Sciences*, 13(2), 129-164.
- Lagnado, D. A. (2011). Causal Thinking. In P. McKay-Illari, F. Russo, & J. Williamson (Eds.), *Causality in the Sciences*. Oxford: Oxford University Press.
- Langley, P. A., & Morecroft, J. D. (2004). Performance and learning in a simulation of oil industry dynamics. *European Journal of Operational Research*, 155(3), 715-732.
- Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth a thousand words. *Cognitive Science*, 11, 65-99.

- Larkin, J., & Simon, H. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11, 65-99.
- Legendre, A. M. (1805). *Nouvelles methodes pour la determination des orbites des cometes*. Paris: Firmin Didot.
- Lehman, D. R., & Nisbett, R. E. (1990). A longitudinal study of the effects of undergraduate training on reasoning. *Developmental Psychology*, 26(6), 952-960.
- Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, 25(3), 265-288.
- Mackie, J. L. (1965). Clauses and conditions. *American Philosophical Quarterly*, 2, 245-264.
- Maldonado, A., Jimenez, G., Herrera, A., Perales, J. C., & Catena, A. (2006). Inattentional blindness for negative relationships in human causal learning. *The Quarterly Journal of Experimental Psychology*, 59(3), 457-470.
- McCrudden, M. T., Schraw, G., & Lehman, S. (2009). The use of adjunct displays to facilitate comprehension of causal relationships in expository text. *Instructional Science*, 37(1), 65-86.
- McCrudden, M. T., Schraw, G., Lehman, S., & Poliquin, A. (2007). The effect of causal diagrams on text learning. *Contemporary Educational Psychology*, 32(3), 367-388.
- Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou, & A. Ortony (Eds.), *Similarity and Analogical reasoning* (pp. 179-196). New York: Cambridge University Press.
- Mill, J. S. (1843/1974). *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence, Methods of Scientific Investigation (Books I-III)* (Vol. VII). (J. M. Robson, Ed.) Toronto: University of Toronto Press.
- Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin*, 117(3), 363-386.
- Pearl, J. (1985). Bayesian networks: A model of self-activated memory for evidential reasoning. *Proceedings of the 7th Conference of the Cognitive Science Society*.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.
- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, MA: Cambridge University Press.

- Petre, M., & Green, T. R. (1993). Learning to read graphics: Some evidence that 'seeing' an information display is an acquired skill. *Journal of Visual Languages & Computing*, 4(1), 55-70.
- Piaget, J. (1930). *The Child's Conception of Physical Causality*. (M. Gabain, Trans.) New York: Harcourt, Brace.
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black, & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.
- Scaife, M., & Rogers, Y. (1996). External cognition: How do graphical representations work? *International Journal of Human-Computer Studies*, 45(2), 185-213.
- Schwanenflugel, P. J., & Shoben, E. J. (1983). Differential context effects in the comprehension of concrete and abstract verbal materials. *Journal of Experimental Psychology*, 9(1), 82-102.
- Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioural and Brain Sciences*, 23, 645-726.
- Stehbens, W. E. (1992). Causality in medical science with particular reference to heart disease and arteriosclerosis. *Perspectives in Biology and Medicine*, 36(1), 97-119.
- Stroop, J. R. (1992). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology: General*, 121(1), 15-23.
- Suwa, M., & Tversky, B. (1997). What do architects and students perceive in their design sketches? A protocol analysis. *Design Studies*, 18, 385-403.
- Tufte, E. R. (1983). *The Visual display of Quantitative Information*. Cheshire, CN: Graphics Press.
- Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.
- Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. In A. Tversky, & D. Kahneman (Eds.), *Judgments Under Uncertainty: Heuristics and Biases* (pp. 117-128). Cambridge, NY: Cambridge University Press.
- Tversky, B. (2005). Prolegomenon to scientific visualizations. In J. K. Gilbert (Ed.), *Visualization in Science Education* (pp. 29-42). Dordrecht: Kluwer.

- Waldmann, J. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning with causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, 121(2), 222-236.
- Wong, P. T., & Weiner, B. (1981). When people ask "why" questions, and the heuristics of attributional search. *Journal of Personality and Social Psychology*, 40(4), 650-663.
- Woodworth, R. S. (1929). *Experimental Psychology*. New York: Holt.
- Wright, S. (1918). On the nature of size factors. *Genetics*, 3, 367-74.
- Wright, S. (1918). On the nature of size factors. *Genetics*, 3, 367-74.
- Wright, S. (1920). The relative importance of heredity and environment in determining the birth weight of guinea pigs. *Proceedings of the National Academy of Sciences*, 6, 320-32.
- Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20, 557-585.
- Yule, G. U. (1899). An investigation into the causes and changes in pauperism in England, chiefly during the last two intercensal decades. *Journal of the Royal Statistical Society*, 62, 249-295.
- Zellner, A. (1988). Causality and causal laws in economics. *Journal of Econometrics*, 39(1-2), 7-21.