International Journal of Artificial Intelligence in Education 19(2009) 425-445
IOS Press

Constructing Causal Diagrams to Learn Deliberation

Matthew W. Easterday, Vincent Aleven, Richard Scheines, Sharon M. Carver,
Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, USA
matteasterday@cmu.edu

Abstract. Policy problems like "What should we do about global warming?" are ill-defined in large part because we do not agree on a system to represent them the way we agree Algebra problems should be represented by equations. As a first step toward building a policy deliberation tutor, we investigated: (a) whether causal diagrams help students learn to evaluate policy options, (b) whether constructing diagrams promotes learning and (c) what difficulties students have constructing and interpreting causal diagrams. The first experiment tested whether providing information as text, text plus a correct diagram, or text plus a diagramming tool helped undergraduates predict the effects of policy options. A second, think-aloud study identified expert and novice errors on the same task. Results showed that constructing and receiving diagrams had different effects on performance and transfer. Students given a correct diagram on a posttest made more correct policy inferences than those given text or a diagramming tool. On a transfer test presented as text only, students who had practiced constructing diagrams made the most correct inferences, even though they did *not* construct diagrams during the transfer test. Qualitative results showed that background knowledge sometimes interfered with diagram interpretation but was also used normatively to augment inferences from the diagram. Taken together, the results suggest that: causal diagrams are a good representation system for a deliberation tutor, tutoring should include diagram construction, and a deliberation tutor must monitor the student's initial beliefs and how they change in response to evidence, perhaps by representing both the evidence provided and the student's synthesized causal model.

Keywords. Ill-defined domains, causal diagrams, external representations, deliberation

INTRODUCTION

It is axiomatic that democracy depends on an active, engaged citizenry–one that can use evidence to reason about policy problems such as: *what should we do about global warming?* or *does junk food advertising lead to childhood obesity?*, in other words a citizenry that can *deliberate*. Deliberation depends on *causal reasoning*, in other words, a citizen must be able to determine if a policy intervention affects a particular outcome before they can balance values, weigh the costs and benefits of different interventions and decide upon the "best" intervention.

Anecdotal evidence suggests that undergraduates have difficulty identifying how interventions affect outcomes. For example, instructors for the service learning class *Technology Consulting in the Community* report that students often struggle to justify how their technology projects will impact the mission of their non-profit clients, so the instructors have begun experimenting with a causal diagramming technique called systems thinking. Similarly, one author found that grant proposals written by novice Peace Corps community organizers often conflate a project's intervention with its purported outcome. Other forms of civic participation at the University such as *Deliberative Polls* (Fishkin, 1995) require similar skills and have started presenting policy information using diagrams. In addition to these examples, there is a more general consensus on the need to teach students to think

critically about policy (Center for Information and Research on Civic Learning and Engagement, 2003) that has not been addressed by AIED research.

As a first step toward identifying a useful representation system for a deliberation tutor, we ask three questions: (a) do causal diagrams improve deliberation compared to text-based representations? (b) if so, do they only provide a perceptual aid to inference, or does constructing a diagram promote a deeper understanding? and (c) what difficulties do students have constructing and interpreting causal diagrams?

For causal diagrams to prove useful, the deliberation task must present some cognitive difficulty, the diagram must make the task easier (e.g., by allowing the student to make inferences perceptually rather than relying on memory) and the student must have acquired the skills to interpret the diagram, and to construct it if necessary. Although it is reasonably clear that deliberation poses a cognitive challenge, it is by no means clear that causal diagrams will improve deliberation or learning. As Ainsworth (2006) notes, while there are many studies showing that diagrams improve reasoning, there are just as many studies showing no benefit, because the usefulness of a diagram depends on the particular task. Classroom studies have shown that diagrams can be more helpful (Pinkwart, Alevan, Ashley, & Lynch 2007; Harrell, 2008; Twardy, 2004), no different (Carr, 2003) or even more difficult (Koedinger and Nathan, 2004) than non-diagrammatic strategies.

Although there is extensive research on the benefit of providing correct diagrams (see Ainsworth, 2006 for an overview, or Mayer, 2001 for work relevant to intelligent tutoring), there are almost no studies on causal diagrams, especially in the realm of policy (see McCrudden, Schraw, Lehman, & Pliquin, 2007 for a recent exception in science).

There is little evidence that constructing diagrams promotes learning, only that it is sometimes a necessary evil for tasks where diagrams are helpful but not provided. Students can have considerable difficulty constructing diagrams (Cox, 1996) and learning to construct diagrams may require extensive training (Grossen & Carnine, 1990). For example, even after two years of instruction, students may not be able to effectively construct equations (Koedinger & Nathan, 2004). While some claim that constructing diagrams promotes deeper understanding, giving students a correct diagram leads to better learning than having them construct one (Stull & Mayer, 2007). Even if students are guided when constructing a diagram, they do not learn significantly more than students who are given a diagram (Hall, Bailey, & Tillman, 1997). In cases where the purpose of construction is to provide a machine readable representation of the student's knowledge for a computer tutor, the high costs of learning construction may simply outweigh the benefits of tutoring.

Assuming that causal diagrams prove useful and that we can isolate the effects of construction and interpretation, we must also identify what difficulties students have learning to use them. To use causal diagrams for policy, students must understand the policy domain, the diagram notation, the mapping between the domain and the diagram, how to make inferences from the diagram, and how to construct the diagram. While each of these poses a potential learning challenge, we do not know a priori where students will have the most difficulty.

The three questions about causal diagrams raised here apply not only to policy, but also to ill-defined domains that rely on causal reasoning such as history (Voss, Carretero, Kennet, & Silfies, 1994), science (Kuhn & Dean, 2004; Zimmerman, 2007; Kuhn, 2005), strategic planning (Huff & Jenkins, 2002), operations research (Narayanan & Armstrong, 2005), medicine (Kuipers & Kassirer, 1984), epidemiology (Joffe & Mindell, 2006) and domains that require representation of conflicting evidence from multiple sources such as argument (Kirschner, Buckingham Shum, & Carr, 2003), intelligence analysis (Heuer, 1999), and legal reasoning (Pinkwart, Aleven, Ashley, & Lynch, 2007).

A Cognitive framework for deliberation

As part of our ongoing research, we have developed a simple task analytic framework for deliberation (Figure 1). The framework provides enough structure to explain the role of causal diagrams in deliberation, to locate points of ill-definition from previous research, and to compare and contrast different tutoring systems. To illustrate the deliberation framework, consider the following example.

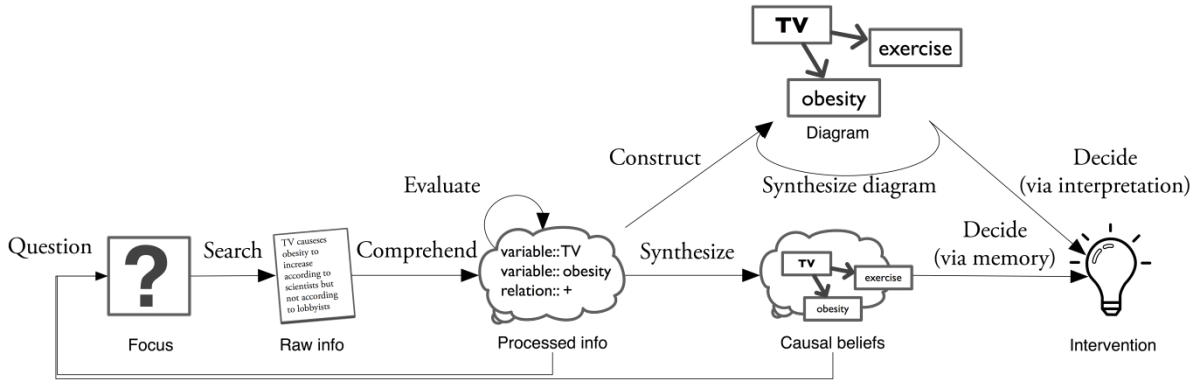


Fig. 1. The deliberation framework roughly defines the stages of problem solving with and without diagrams.

The citizen must first begin with a **question** such as: "what should we do about childhood obesity?" or "should we limit junk food advertising on television?" As in other ill-defined domains, the initial question might be considerably vague and require additional effort to define (Rittle & Webber, 1973; Voss, 2005; Lynch, Ashley, Aleven & Pinkwart, 2006; Simon, 1996).

With the question in hand, the citizen must then **search** for relevant information. She might consult common knowledge to recall that "exercise decreases obesity," search the internet for scientific reports about the effects of junk food advertising, elicit information from a third party, or if she has additional expertise in research, conduct experiments and observational studies. Because policy problems are ill-defined, the search space is typically larger than the citizen can fully search, and in some cases the information needed to definitively solve the problem may not exist (Voss, 2005; Rittle & Webber, 1973; Simon, 1996; Horn & Webber, 2007).

After acquiring a piece of raw information, such as a report on the effects of junk food advertising on childhood obesity, she must **comprehend** the relevant information in the article. For example, she might identify junk food advertising and childhood obesity as variables, the causal relations among the variables, (e.g., that advertising increases obesity), the source making the claim, (e.g., Dr. Neuringer from Johns Hopkins University), and the type of information, (e.g., an experiment).

The outcome of this comprehension process is some schematized mental representation. The citizen should ideally **evaluate** the strength of the information at this point, for example recognizing the Johns Hopkins clinical trial as a stronger piece of evidence than a claim from Aunt Louise. There is no normative theory for evaluating evidence, that is, how much an observational study is *worth* compared to an experiment or to a mechanistic explanation, and impartial evidence evaluation of policy information proves difficult (Taber & Lodge, 2006, Lord, Ross, & Lepper 1979, Kuhn, Amsel, & O'Loughlin, 1988).

After comprehending and evaluating each new claim, the citizen must **synthesize** this information with his other beliefs. If the citizen has no prior beliefs about the effect of advertising on obesity, he might simply accept the evidence at face value that junk food commercials have a deleterious effect on obesity. On the other hand, the citizen might believe that junk food commercials don't affect obesity based on some other evidence, perhaps other experimental studies showing no effect of advertising on obesity. In this case, the citizen should acknowledge the study, perhaps by lowering his confidence in his original belief, but may ultimately overrule this particular piece of information. The ill-definition in evaluation propagates to synthesis. If two pieces of evidence contradict each other, what should the citizen conclude? There are some normative constraints on synthesis but, again, no well-defined algorithm.

Through this process of search and analysis, the citizen builds some causal model of the evidence (Jones & Read, 2005) encompassing all the discovered claims and evidence relevant to the policy problem including: common knowledge that exercise and junk food affect obesity, scientific reports from experts that watching TV does not affect the amount children exercise, conflicting unresolved claims such as that ads do increase obesity according to an advocacy group, but that junk food commercials only affect the brand eaten according to junk food lobbyists, and so on (see Britt, Rouet, Georgi, & Perfetti, 1994, and Perfetti, Rouet, & Britt, 1999 for empirical and theoretical accounts of representing causal models of evidence in history, and Chinn & Brewer, 2001 for causal models of evidence in science). The variability in the earlier steps of search, evaluation and synthesis may lead citizens to create different, yet plausible, models of the same problem, leading to the problem of *multiple representations* often seen in ill-defined domains (Horn & Webber, 2007; Voss, 2005; Lynch et al., 2006; Simon, 1996), a point to which we will return.

Finally, with this synthesized model of the policy domain, the citizen is now in a position to **decide** upon a policy recommendation (comparing alternatives in the policy literature, e.g., Patton & Sawicki, 1993; Walker & Fisher, 1994). The citizen must take into account different possible interventions, (e.g., limiting junk food advertising, starting school exercise programs), different possible outcomes (e.g., decreasing obesity and cost), and the desirability of different outcomes to different stakeholders. If the citizen can find a policy intervention that satisfies all these constraints, then she is ready to make a recommendation. If not, she may have to redefine the question, search for more information, or simply identify the least objectionable policy. Even at this point, when the citizen is balancing the values of different stakeholders, her underlying causal reasoning must be sound to make these tradeoffs effectively. Even here, causal reasoning is essential. Given the variability in problem solving noted earlier, one can see that even if two citizens were to use the same decision-making procedure, they might still reach different conclusions, hence ill-defined problems like these are thought to lack a single correct answer (Lynch et al., 2006; Voss, 2005; Horn & Webber, 2007; Rittle & Webber, 1973).

Up to this point, the framework has described reasoning as if it takes place entirely within the citizen's head, without any external representations or tools. Reasoning about policy in this way would be like solving Algebra problems without writing equations. The framework conjectures that an appropriate diagrammatic representation (with sufficient training) will improve deliberation in the same way that equations improve Algebraic problem solving. To understand how external representations such as causal diagrams affect reasoning, let's reconsider the previous example at the point where a citizen has acquired a new piece of information.

Once raw information such as a scientific report about the effects of junk food advertising has been comprehended, the next step is to **construct** a representation of that information. For example, if

the report says that advertising increases the amount of junk food eaten, the citizen could construct a diagrammatic element like that in Figure 2 (left).

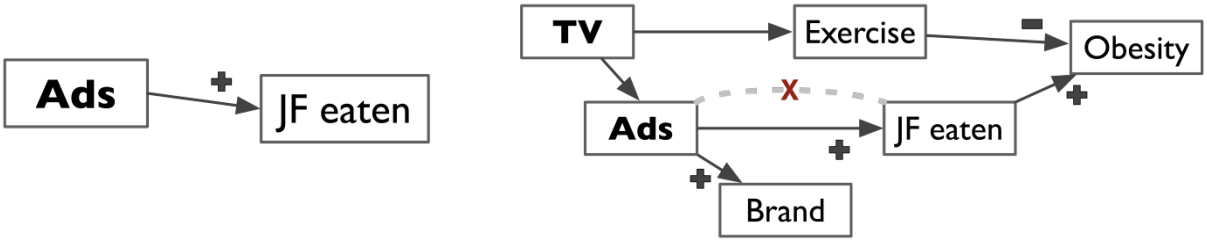


Fig. 2. A diagram element representing the causal claim that advertising increases junk food eaten (left) and a whole causal model of advertising and obesity (right).

Each time the citizen encounters a piece of information, he must update his diagrammatic representation. Over time, through this process of diagram construction, the citizen builds a representation of the policy domain like that in Figure 2 (right).

In the early phases of problem solving, the citizen created a diagrammatic representation of the problem that she must now **interpret**, balancing values, weighing the costs and benefits of different interventions and deciding upon the "best" intervention. These tasks correspond to inferences made from the causal diagram, such as identifying: which outcome variables of the diagram are of interest to different stakeholders, the resources needed to manipulate targeted variables, the tradeoffs associated with positive and negative causal impacts of the targeted variables on the outcome variables, and so on (see Montibeller & Belton, 2006 on the role of causal diagrams in decision).

By delineating the steps of questioning, search, comprehension, and evaluation, the steps of synthesis, and decision along the standard path, and the steps of construction and interpretation along the diagram path, the deliberation framework provides us with a rough cognitive model for deliberation, explains the role of causal diagrams in deliberation, locates well-known characteristics of ill-definition at specific points in the reasoning process and will allow us to compare and contrast intelligent tutoring research across domains.

Causal diagrams and the problem of multiple representations in ill-defined domains

As described in the deliberation framework, during synthesis we confront the problem of *multiple representations,* i.e., that there is no agreed upon representation of an ill-defined problem. Different researchers have described this in several ways: Horn and Webber (2007) note that there is no correct view of the problem, Simon (1996) describes the challenges of representing social planning problems, Rittle & Webber (1973) argue that there is no definitive formulation of the problem, and Lynch et. al. (2006) point out that one has to represent *open-textured concepts.* The key point is that two citizens working on the same problem may produce two different representations that are both "correct."

We can recast the problem of multiple representations more precisely in terms of the deliberation framework: two citizens might produce different representations either because they *select* a different *representation system* or because they *construct* different particular representations within the representation system. For example, two citizens might select different representation systems if one uses causal diagrams and another uses argument diagrams. If both citizens were to select causal diagrams as the representation system, they still might *construct* different particular representations if

one creates the diagram in Figure 2, while another creates a diagram with different variables, say removing the *brand* variable. In well-defined domains like Algebra, there is consensus about both the representation system to select and the particular representation to construct.

For deliberation, there is no consensus about which representation system to select. Besides causal diagrams, there are also argument diagrams (van Gelder, 2003), concept maps (Kirschner et al., 2003), evidence maps (Suthers, Weiner, Connelly, & Paolucci, 1995), or no representation (other than text) at all. Although causal diagrams are not the only possible representation system for deliberation, we chose to investigate causal diagrams because of the centrality of causal reasoning in policy (Pawson, 2006), the widespread use of causal diagrams in strategic planning (Huff & Jenkins, 2002; Narayanan & Armstrong, 2005) the tendency of political experts to solve policy problems using a causal strategy (Voss, Tyler, & Yengo, 1983), the two decades of research on formalizing causal graphs (Spirtes, Glymour, & Scheines, 2000; Pearl, 2000), the machine readability of causal graphs even in their qualitative form, and the fact that most causal reasoning tutors use causal diagrams or text.

By testing which representation systems best improve deliberation, we may eventually reach consensus on how to teach deliberation. Achieving this goal would add significant definition to the domain because it would identify the types of information to search for, what to represent during synthesis, and potentially formalize inferences made during decision. Although we cannot test all representation systems in one paper, comparing causal diagrams to text-based representations advances this goal.

Intelligent tutoring systems and causal reasoning

Turning to instruction, there are several intelligent tutors that use causal reasoning problems, which we can analyze using the deliberation framework. In Betty's Brain, students construct causal diagrams on global warming which represent the knowledge of a virtual student named Betty (Leelawong & Biswas, 2008). Betty passes or fails her quizzes based on the accuracy of the diagram. In 20/20, students build causal diagrams of the English Civil war by selecting causes from a list (Masterman, 2005). VModel allows students to construct causal diagrams and run simulations (Forbus, Carney, Sherin, & Ureel, 2005). Students using SEEK (Graesser, Wiley, Goldman, O'Reilly, Jeon, & McDaniel, 2007) and Sourcer's Apprentice (Britt & Aglinskas, 2002) read evidence of varying reliability before writing essays on the causes of volcanic activity or the Panamanian revolution.

With respect to deliberation, only Betty's Brain and Sourcer's apprentice allow students to search for information. Betty's Brain provides hyperlinked pages containing only accurate information and Sourcer's Apprentice provides a set of seven sources, but in neither case is search a focus of tutoring. SEEK requires students to read all information whereas 20/20 and VModel do not provide information as part of the tutoring environment. Note that tutors *can* teach search using microworlds (Jonassen & Ionas, 2008) like the Causality Lab (Scheines, Easterday, & Danks, 2007) in which students collect arbitrary amounts of experimental data.

To teach comprehension and evaluation, SEEK and Sourcer's Apprentice present evidence of varying reliability and provide students with a set of forms to help them think critically about the source and the source's causal claims. Sourcer's Apprentice emphasizes comprehension, requiring students to explicitly select text containing source information. SEEK emphasizes evaluation, providing feedback on the student's ratings of the source's reliability. In contrast, Betty's Brain provides only reliable information, so no evaluation tutoring is provided.

The tutors that use text-based representations (SEEK and Sourcer's Apprentice) do not support synthesis other than through the structured notes created during comprehension and evaluation.

Tutors using causal diagrams provide feedback on construction in different ways. 20/20 provides the most explicit feedback by immediately comparing students' diagrams with an expert model. In Betty's Brain, Betty fails her quiz if the diagram is incorrect, indicating that the diagram does not match the expert model. VModel provides general construction feedback when the student makes syntactical errors, such as making a causal arrow between boxes that represent an entity rather than two boxes representing a quantitative parameter.

Only Betty's Brain and VModel provide feedback on diagram interpretation. Betty's Brain provides feedback on interpretation when the student asks Betty to predict and explain the effect of one variable on another. In VModel, the student can test his prediction by running a simulation. Modeling tools like VisiGarp (Salles, Bredeweg, & Araújo, 2006) can also provide feedback in this way. 20/20 provides no feedback on interpretation because the causal diagram itself is the answer.

The deliberation framework shows the different ways tutors address the problem of multiple representations (Table 1). SEEK and Sourcer's Apprentice do support synthesis, i.e., they essentially ignore the problem of multiple representations. VModel relies on human tutors to provide feedback about the accuracy of the diagram. Betty's Brain and 20/20 provide feedback on how closely the student's model conforms to an expert model. These differences may result in part from the types of information provided. SEEK and Sourcer's Apprentice present conflicting evidence and help the students to represent pieces of evidence in isolation. Betty's Brain, 20/20 and VModel allow representation of a single "true" model, i.e., they do not try to represent conflicting evidence. To target the challenge of ill-defined problems explicitly, this project uses causal diagrams to represent the relations between conflicting evidence.

Table 1
Scaffolding and Feedback (Fdbk) Provided by Causal Reasoning Tutors on Steps of Deliberation

	Tutor				
Step	Betty's Brain	VModel	20/20	SEEK	Sourcer's Apprent.
Question	-	-	-	-	-
Search	Reliable	-	-	Vary reliability	Vary reliability
Comprehend	-	-	-	Form	Form + fdbk
Evaluate	-	-	-	Credibility fdbk	-
Text path					
Synthesize	-	-	-	-	-
Decide	-	-	-	-	-
Diagram path					
Construct	Betty's quizzes	Syntactic fdbk	Model fdbk	-	-
Interpret	Betty's explanations	Predictions fdbk	-	-	-
Tests	Content knowledge	Modeling skill	Model	Sourcing skill	Sourcing skill

The deliberation framework also shows that the design decisions of these tutors are not fully supported by the scant research on causal diagrams. The research on the benefits of diagrams suggests that SEEK and Sourcer's Apprentice could improve synthesis (not just sourcing) by using diagrams, but the argument for using causal diagrams is not strongly supported. If the purpose of Betty's Brain

is to teach content knowledge, then the general research on diagrams suggests it may be better to provide a correct diagram and have students help Betty interpret it, rather than to have students construct the diagram themselves. 20/20 tests the accuracy of the constructed diagram when supported by tutoring, but the research suggests that students will probably not be able to construct these diagrams when the tutor is removed. Different causal reasoning tutors have made different choices about whether to use causal diagrams, and all of those using diagrams have made a risky decision to emphasize construction. Given the lack of empirical research validating the design of causal reasoning tutors, our project focused on empirically testing the effects of causal diagram interpretation and construction on performance and learning.

Research question: Constructing and interpreting causal diagrams to learn deliberation

To determine: (a) whether causal diagrams improve deliberation compared to text, (b) whether construction promotes learning, and (c) the learning difficulties associated with constructing and interpreting causal diagrams, we examined the effect of causal diagrams on students' policy recommendations given evidence from conflicting sources. To isolate the processes of text-based synthesis, diagram construction, and diagram interpretation (see Figure 1), our study used three levels of external representation: text only, in which students solve problems unguided by an external representation, text plus a diagram tool, in which students solved the problem using a causal diagram, and text plus a correct diagram, in which students solved the problem using a causal diagram, but bypassed the process of construction. These levels correspond to three competing hypotheses:

1. *Text hypothesis.* Neither reading nor constructing causal diagrams will improve performance because the costs of learning to construct and interpret diagrams outweigh any benefit that diagrams might provide relative to text.
2. *Diagram hypothesis.* Having a correct causal diagram will improve performance because the diagram bypasses the process of synthesizing a causal model with an easier perceptual process of diagram interpretation, and also avoids the errors and extra burden of diagram construction.
3. *Tool hypothesis.* Constructing causal diagrams will improve performance because constructing a diagram forces one to more deeply process information.

Because we are interested both in deliberative tasks where citizens may be provided with diagrams, such as deliberative polls, and tasks where no diagram is provided, such as community organizing, we tested students on a posttest where diagrams or tools were provided and a transfer test where only text was provided. This also allowed us to separate the effects on learning and performance.

The task asked students to predict the effect of a policy intervention on a given outcome assuming a given set of sources were credible. This task isolates the effects of text-based and diagram-based synthesis from other stages of processing such as search, because errors in other stages would mask the differences between text and diagrams on synthesis. Likewise, the task tried to minimize comprehension errors by using short texts with clearly named policy variables. To minimize variability in students' evaluation of sources, the task controlled evaluation by asking students to assume a given set of sources were credible. The nature of the task and the procedure students learned for interpreting evidence meant that questions had single, correct answers. Restricting the task in this way made it less ill-defined, however, the strategy was to establish whether there was *any* benefit of causal diagrams during the stages of deliberation where they should have the greatest influence on reasoning. If diagrams indeed provide benefit, then future work will investigate more complicated, ill-defined tasks (Easterday, Aleven, Scheines, & Carver, 2009).

The studies described herein both used the same three group between-subjects design and were presented on-line (Easterday, Aleven, & Scheines, 2007a, 2007b).

EXPERIMENT 1: TEXT, DIAGRAMS AND TOOLS

Method

Participants. 64 University students who had no prior training in causal reasoning were recruited through introductory philosophy classes and campus flyers and paid $10 for their time. One student who did not complete the study due to technical difficulties was dropped from the study. The remaining 63 students were 57% male and 43% female, with a mean age of 21 years (SD=3.36). The majority of students were born in the U.S. (86%), and were native English speakers (92%). All students reported using the internet at least once a day, and all but one student reported using the internet several times a day. Students were 56% Caucasian, 15% Asian/Pacific Islander, 3% African American, with 22% of students declining to identify, and 3% not identifying with a specific category.

Procedure. Students were randomly assigned with replacement to either the text, diagram, or tool groups, then completed a pretest given in text, followed by a short training, then a posttest given either in text, text plus a correct diagram, or text plus a diagram tool, and finally a near transfer test given in text only (Figure 3).

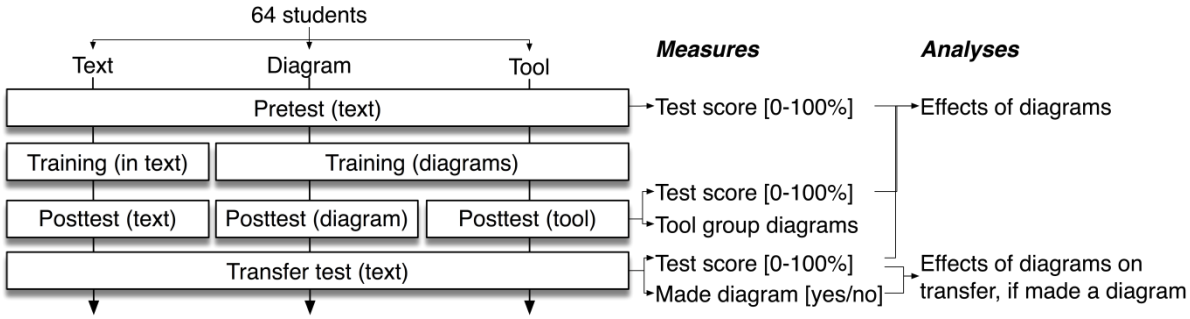


Fig. 3. Experimental procedure, measures and analyses.

The **pretest** consisted of a 234 word text on global warming, in which human activity affected species loss through habitat destruction according to common knowledge and through increased carbon dioxide according to some sources, or only through natural geological change according to other sources. Students answered 10 questions about how intervening on one variable would affect another variable, according to different sets of sources. The causal model in the pretest had a structure identical to that in the posttest (Figure 4) and transfer test, and questions on the same causal relations.

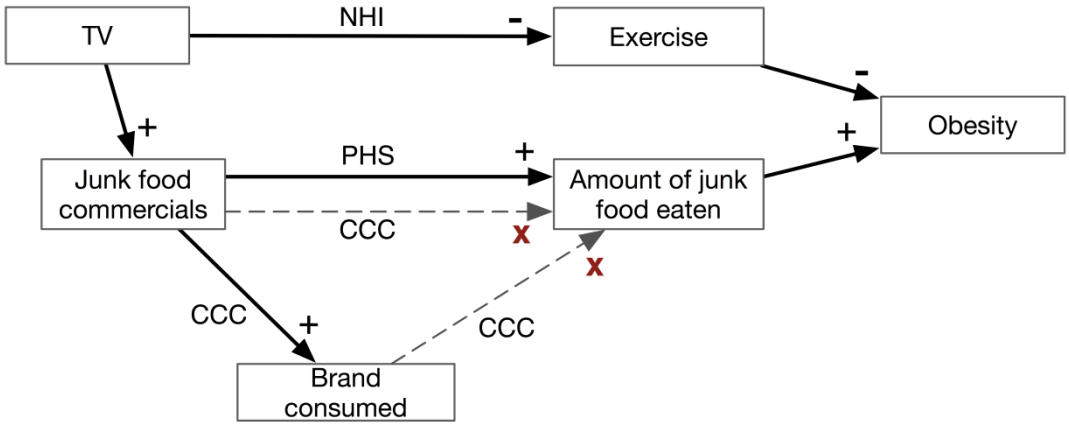
In the first training exercise, students were given a 63 word paragraph about smoking, where according to common knowledge, smoking causes stained teeth and lung cancer causes early death, but in which researchers and tobacco companies disagree about whether smoking causes lung cancer. The students then answered 9 questions about causation and correlation similar to those on the pretest, such as, "According to the NHI does smoking increase your chances of getting lung cancer?" Students received feedback with explanations immediately after each answer.

Childhood obesity is now a major national health epidemic. A number of facts are widely agreed upon by the public and scientific community: doing exercise decreases obesity, and eating junk food increases obesity. It's also clear that people who watch more TV are exposed to more junk food commercials.

Parents for Healthy Schools (PHS), an advocacy group which fought successfully to remove vending machines from Northern Californian schools, claims that junk-food commercials on children's television programming have a definite effect on the amount of junk food children eat. In a recent press conference, Susan Watters, the president of PHS stated that "...if the food companies aren't willing to act responsibly, then the parents need to fight to get junk food advertising off the air."

A prominent Washington lobbyist Samuel Berman, who runs the Center for Consumer Choice (CCC), a nonprofit advocacy group financed by the food and restaurant industries, argues that junk food commercials only "influence the brand of food consumers choose and do not not affect the amount of food consumed." While Mr. Berman acknowledges that watching more TV may cause people to see more junk food commercials, he remains strongly opposed to any governmental regulation of food product advertising.

Recent studies by scientists at the National Health Institute have shown that watching more TV does cause people to exercise less.



1. According to the NHI, will making children exercise more reduce childhood obesity?
2. According the the NHI and CCC, will making children watch less TV decrease childhood obesity?
3. According to the CCC and PHS, will reducing the number of junk food commercials children watch reduce childhood obesity?
4. According to the CCC and PHS, will reducing the number of junk food commercials children watch reduce the amount of junk food they eat?
5. According to the PHS, will watching TV cause children to exercise less?
6. According to common knowledge, will making children watch less TV decrease childhood obesity?
7. According to the NHI, will making kids exercise more reduce the number of junk food commercials they watch?
8. According to the NHI, will reducing the number of junk food commercials children watch reduce childhood obesity?
9. According to common knowledge, will making kids exercise more reduce the amount of junk food they eat?
10. According to the PHS, will making kids exercise more reduce the number of junk food commercials they watch?

Fig. 4. Policy text, correct diagram (diagram group only), and questions on the posttest.

In the second training exercise, students received direct instruction providing detailed answers to four **questions** in the first training exercise illustrating the causal model of the researchers, the causal model of the tobacco company, conflicts between the two models, and the difference between causation and correlation. For the diagram and tool groups, instruction was presented in diagrammatic representations of the claims; for the text group, with relevant claims highlighted in the text.

In the third training exercise, students answered six questions about the global warming testimony from the pretest, but this time each question was answered in five steps each of which included correctness feedback and an explanation. In the first step, students identified the variables in the question. In the second step, students identified whether the question was about cause or correlation. In the third step, students identified the sources in the question. In the fourth step, if the question was causal, students identified whether there was a causal chain from the first variable to the second, no causal chain, or a chain according to one source but not another. If the question was correlational, students identified whether there was a common cause, no common cause, or a common cause according to one source but not another. In the final step, students answered the original question, either a causal question of the form: "according to the sources would the intervention affect the outcome?" or a correlation question of the form: "according to the sources would the two variables be associated?" Students had to answer each step correctly before proceeding to the next step.

In the fourth training exercise, given only to students in the diagram and tool groups, students reproduced a simple 4 variable diagram using the diagram tool. Students were not given feedback.

In the fifth training exercise, students were given a 108 word version of the pretest text. Diagram and tool students were asked to "try constructing a diagram for the testimony" and text students to "try to extract and summarize the causal information for the testimony." At any time, students could click a "show answer" button to see an expert solution. For the diagram and tool students, the answer included a causal diagram, but for the text students, a bulleted list of causal claims.

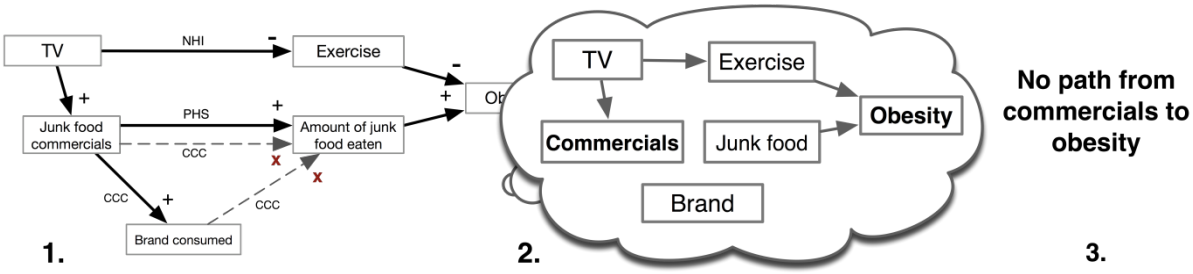


Fig. 5. To answer the question: "According to the NHI, will reducing the number of junk food commercials children watch reduce childhood obesity?" the student takes the claims of the NHI and the claims of common knowledge (unlabeled arrows) from the original diagram (1) resulting in the diagram (2). Then the student looks for a path between commercials and obesity, and finding no link (3), answers no.

The **posttest** (Figure 4) consisted of a 223 word text on junk food advertising and childhood obesity with the same causal structure as the pretest. The text group received only the text, while the diagram group received the text with a correct diagram, and the tool group received the text and a diagramming tool (Easterday, Kanarek & Harrell, 2009) with which they could construct their own diagram. Students were then asked 10 multiple-choice questions of 5 different types: *chain* questions (1&2) with a causal chain from the first variable in the question to the second according to the sources in the question, for which the correct answer is yes, *conflict* questions (3&4) where the sources disagree about the causal path, for which the correct answer is "inconclusive," *no path* questions

(5&6) where there is no causal path, so the correct answer is "no," and *common cause* (7&8) and *common effect* questions (9&10) in which a third variable either causes, or is caused by the variables in the question, for which the correct answer is no.

The (near) **transfer test** consisted of a 201 word text on the deterrent affect of *three strikes laws* on crime that had the same causal structure as the pretest and posttest. Like the pretest, all students received the description as text only, although they could take notes or draw diagrams on scratch paper which was later collected. Students answered 10 questions like those in the pretest and posttest.

Results

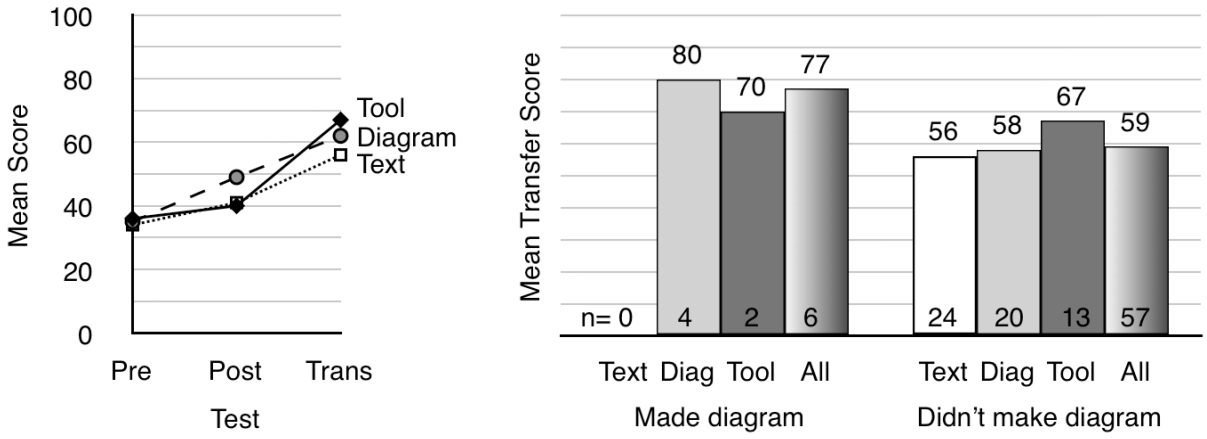


Fig. 6. Test scores for text, diagram, and tool students (left) and transfer test scores for students who made, or did not make diagrams (right).

Effects of diagrams. The pretest, posttest, and transfer test scores of each group are shown in Figure 6 (left). On the **pretest**, The text (n= 24, M=34%, SD=11), diagram (n = 24, M=35%, SD=11) and tool (n=15, M=36%, SD=17) groups all performed at chance. A linear regression analysis showed no significant effect of condition on pretest scores, $F(2,60) = 0.145$, $p > .86$.

After training, on the **posttest** where students were given policy information either as as text, text with a correct diagram, or as text with a diagramming tool, diagram students scored higher (M=49%, SD= 26) than text students (M=41%, SD=23) and tool students (M=40%, SD=22). Posttest scores were regressed on condition, time on training and time on posttest. These three predictors accounted for 30% of the variance in posttest scores, which was highly significant, $F(8,54)=4.30$, $p < .0005$. Having a correct diagram ($b=41$, $p < .04$) and spending a longer time on training ($b=3.2$, $p < .05$) both significantly increased posttest scores. Two interactions showed that for students who had a correct diagram, there was an increase in posttest scores for those who spent a *shorter time* on training ($b=-4.8$, $p < .01$), and those who spent a longer time on test ($b=6.9$, $p < .03$).

Despite the superior performance of the diagram students on the posttest, on the **transfer test** in which all students received policy information as text only, tool students (M=67%, SD=15) had higher scores than both diagram students (M=62%, SD=20) and text students (M=56%, SD=22%). Regressing transfer test scores on condition and posttest score showed that these two predictors accounted for 36% of the variance which was highly significant $F(3,59)=6.47$, $p < .0000015$. Students in the tool condition had significantly higher transfer test scores than students in the text condition

(b=12, $p < .03$), and students who had higher posttest scores also had significantly higher transfer test scores (b=5.0, $p < .0000002$). Transfer test scores of diagram and tool students were not significantly different (b=1.7, $p > .71$). A comparison of the tool and diagram students showed that tool students scored significantly higher than diagram students (b=9.8, $p < .04$), $F(2,36)=13.85$, $p < .00003$.

Diagrams constructed by tool group on the posttest. Unfortunately, the diagram construction log data for 7 of the 15 tool students was corrupted. Logs for the remaining eight students showed that they all made diagrams and that no student made a perfect diagram. The best diagram contained all variables except *brand consumed* and six out of eight causal arrows, all of which were correctly labeled. An intermediate diagram contained all variables except *brand consumed* and five causal arrows, none of which were labeled. The worst diagram had seven boxes, five of which contained entire causal claims from the text and two of which contained other sentences from the text; the principle by which arrows connected the boxes was unclear.

Effects of diagrams on transfer conditional on making a diagram. To better understand the tool group's learning gains, we looked separately at students who made or did not make a diagram on scratch paper during the transfer test (in Figure 6 right, compare *all* students who *made diagrams* to *all* students who *did not make diagrams*). The six students who made diagrams on scratch paper had higher transfer test scores (M=77%, SD=14) than the 57 students who did not make diagrams (M=59%, SD=20). A regression analysis showed that making a diagram was a significant predictor of transfer scores (b=17, $p<0.04$), accounting for 5% of the variance, $F(1,61)=4.28$, $p < .04$. The higher transfer test scores of students who made diagrams suggest either that diagrams are useful or a selection effect where the "good" students made diagrams.

The transfer scores of the six students who made diagrams on scratch paper during the transfer test (only two of which are from the tool group) cannot account for the higher transfer scores of the tool group as a whole. The tool group had higher transfer scores because, among the vast majority of students who did not make diagrams (13 in the tool group, 20 in the diagram group, 24 in the text group), the tool students who did not make diagrams had higher transfer test scores (M=67%, SD=16) than the diagram (M=58%, SD=19) and text (M=56%, SD=22) students who did not make diagrams. Regressing the transfer test scores of students who did not make diagrams on condition and time on the transfer test showed that these two predictors accounted for 20% of the variance, $F(3,53)=5.75$, $p < .002$. Students in the tool condition had significantly higher transfer test scores (b=14, $p < 0.02$) than students in the text condition, and students who spent longer on the transfer test (b=5.3, $p < 0.0004$) also had significantly higher transfer scores. Although these results are correlational, they suggest that, when diagrams are unavailable, having practiced constructing diagrams (on the posttest) led to higher scores on the transfer test even if one did not make a diagram.

Time. There were no significant differences in time between groups on the pretest, posttest or transfer test. When controlled for the time that the diagram and tool groups spent learning the tool buttons on the fourth training exercise (M=1.6 min, SD=1.8), and gender, we find no significant difference in training time.

Discussion

The purpose of this study was to determine: (a) whether causal diagrams improve deliberation compared to text and (b) whether construction promotes learning. The results of the posttest suggest that causal diagrams do indeed provide a good representation system for deliberation. Furthermore, even if students cannot, or will not, make diagrams on the transfer test, the act of having practiced constructing diagrams improves future deliberation. It is possible that the benefit of construction

practice arises because diagram construction forces students to explicitly identify variables and causal relations (i.e., to practice *comprehension*) a skill that can be used even when not using diagrams.

STUDY 2: EXPERT / NOVICE THINK-ALOUDS

Method

Participants. To gain a better understanding of the types of errors students make when reading and constructing diagrams, this study compared students' performance with the performance of causal reasoning experts. Participants included 4 undergrad *novices* and 3 faculty and graduate student *experts* who all had doctorate degrees in philosophy and had conducted original research on causal reasoning. All participants were offered $20 but all experts declined payment.

Procedure. The procedure was identical to the first study except that participants were asked to think-aloud while a screen capture program recorded their speech and on-screen behavior. Because the long term goal is to develop a deliberation tutor, we did not try to quantify the frequency of these errors, but informally identified the *types* of errors to later develop measures that will allow a deliberation tutor to detect and respond to these errors.

Results

Using Text (Novice 1, Expert 1). Both the novice and expert in the text condition performed quite poorly. Novice 1 scored 20% while Expert 1 scored 0% on the first half of the questions, after which Expert 1 ended the experiment stating, "my brain is fried." While Expert 1's performance seems abysmal, recall that he was not allowed to use his standard tool (a causal diagram). Unlike Novice 1, Expert 1 realized the difficulty of completing the task without a diagram. This performance underscores the difficulty of reasoning about even simple causal systems using text alone.

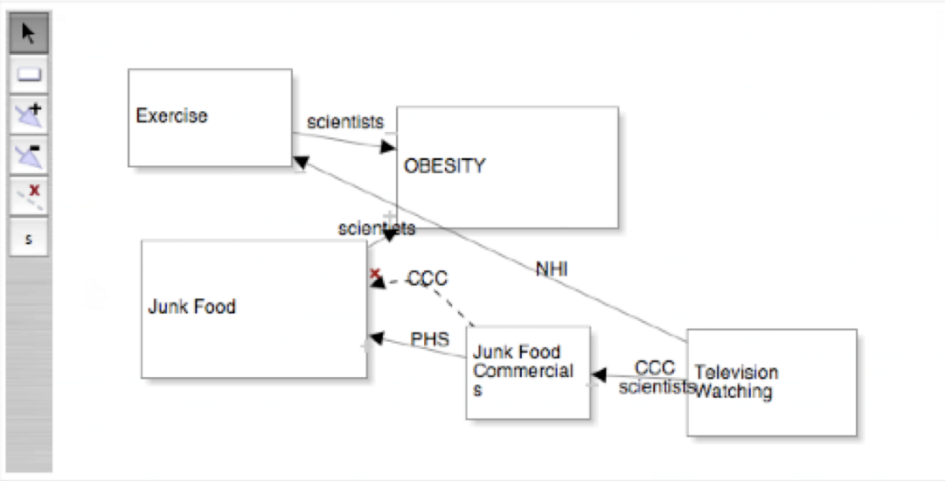


Fig. 7. Novice 2's diagram omits the "brand" variable and mislabeled links from junk food to exercise.

Constructing diagrams (Novice 2, Expert 2). In the first study, students who were given case studies as text accompanied by a diagramming tool scored an average of 40%, performing no better than the text group. Given the poor performance of this diagram construction group in the previous study, we expected Novice 2 to have difficulty with diagram construction. In fact, both Novice 2 and Expert 2 made better diagrams than most observed in the previous study. Despite making relatively good diagrams, small errors in diagram construction sometimes lead to relatively large errors in interpretation. For example, by mislabeling the two arrows pointing to obesity (Figure 7), Novice 2 might answer every question on the posttest incorrectly. While both their diagrams contained errors, Expert 2's diagram (assuming it was used correctly to answer the test questions) would have led to the correct answer on 100% of the questions, whereas Novice 2's diagram would have led to the correct answer on 20% of the questions.

Interpreting diagrams (Novice 3 & 4, Expert 3). Participants' background knowledge and beliefs often interfered with their interpretation of the diagram (Table 2).

The first two types of errors: *override* and *speculation* occurred because participants claimed relevant knowledge not described by the diagram. In the ***override error***, the reasoner correctly reads the diagram, but decides that his background knowledge is more credible. For example, on question 10, Expert 3 correctly interpreted the diagram, stated that this conclusion contradicts his background knowledge, and then decided that his background knowledge was superior.

> Naturally I would assume that the PHS people would say "yeah it will reduce the number of junk food commercials they watch" because in fact, this guy up here, I think most people would think is actually a, uh, uh, goes both ways.… However, I'm supposed to answer the question based on what's been given to me so far... So I'm going to say the answer I'm supposed to give is 'no', but quite frankly, well you know what, I'm going to give the answer I think is right given the sorts of things I've got here, which is that it's actually inconclusive.

This error can be normative if the participant makes separate and correct predictions about the both the evidence provided and his beliefs and can show that his belief is more credible than the evidence provided. In the ***speculation error,*** the participant adds information to the diagram about what a source would say, given what that source has already said. On question 5, Expert 3 speculated that the PHS would accept the NHI's claim that *TV* affects *exercise*:

> Well I'm willing to bet the PHS would absorb… well it's inconclusive, we don't know what the PHS thinks, we aren't given any context. ...So I'm going to say inconclusive, because I was not given that piece of information. Moreover, I think the PHS would presumably accept those kind of studies.

Because *override* and *speculation* errors are caused by background knowledge not represented in the diagram, they can be thought of errors in the construction phase of deliberation, rather than interpretation–it's not that the participant incorrectly interprets the diagram so much as the diagram doesn't represent all the information being used to solve the problem.

The third and fourth types of errors: *reverse causation* and *false uncertainty* errors also result from background beliefs but in a non-normative way when the subject misinterprets the meaning of an arrow to produce an interpretation consistent with his beliefs. In a ***reverse causation error***, the participant selectively interprets an arrow indicating that A causes B to also indicate that B causes A. Novice 3 and 4 both made reverse causation errors on question 7 when they reinterpreted an arrow showing that watching TV decreases exercise to also mean that increasing exercise will decrease TV watching. Novice 4 says:

> Well without looking at that I would say 'yes', but looking at this…so kids are exercising more, then they watch less TV, which means they have, watch less junk food commercials. But the question is…'will

> making children exercise more, reduce the number of commercials they watch'. I don't know about reading the graph backwards, it's confusing. Well I'm going to say 'yes'.

Again, Novice 4 did not systematically interpret arrows this way on other questions, but only when such a reinterpretation rendered the diagram consistent with her background knowledge. In a **false uncertainty error,** the participant selectively interprets the lack of an arrow by a source as indicating that "we don't know what the source thinks" instead of that "the source makes no claim" as was taught during training. For example, on question 8 which asks about the NHI (and common knowledge), neither the NHI nor common knowledge make any claims about the effect of junk food commercials on the amount of junk food eaten, which according to the rules taught in the training means that the NHI does not think JF commercials affect junk food eaten. However, Novice 4 says: "it doesn't say anything on here... I can't tell from there, so from looking at that, that would be inconclusive..." and on question 6: "it doesn't say anything about junk food commercials, so that would be inconclusive," which are incorrect answers. If Novice 4 just misunderstood how to interpret lack of an arrow, then she would not have answered question 5 correctly. Later, we see that she can infer the correct answer of "no" but overrides this answer because it contradicts her background knowledge. On question 6, in which there is no path from TV to obesity through either exercise or junk food eaten, Novice 4 says: "I would assume that if you're watching TV you're not playing…that would lead to less children being obese." The quote suggests that Novice 4 wanted to answer *yes* according to her background knowledge, and *selectively* reinterpreted the meaning of an absence of an arrow when the correct interpretation contradicted her belief. When asked why she chose *inconclusive* rather than *no*, she responded: "…my feeling is to go for *yes*, so I kind of compromised and went for inconclusive" indicating that indeed background knowledge is selectively influencing her interpretation of the diagram.

The last two types of errors, *chaining* and *impasse* result simply from being unable to combine the diagrammatic elements to make the proper inference. In a ***chaining error***, the participant notices the relevant arrows but does not combine them correctly to make the proper inference. In an ***impasse error***, the participant simply gives up on the diagram (and text) altogether.

Table 2
Errors by Participants in the Diagram Condition

	Error		
Question	Novice 3	Novice 4	Expert 3
1	+	+	+
2	Chaining	+	Chaining
3	+	+	+
4	+	+	+
5	+	+	False uncertainty, Speculation
6	+	False uncertainty	+
7	Reverse causation	Reverse causation	+
8	+	False uncertainty	False uncertainty
9	+	Impasse	+
10	+	Impasse	Override
% correct	80	50	60

Note. Cells with a "+" indicate the participant answered the question correctly.

Discussion

The purpose of the second study was to identify the learning difficulties associated with constructing and interpreting causal diagrams in order to develop a deliberation tutor that can detect and respond to these errors. Results identified several types of errors that arise during the construction and interpretation phases of deliberation. Errors in diagram construction can reflect upstream errors in comprehension (as when a novice misses a claim) and from background knowledge not present in the diagram that might be used to solve the problem. Even with decent performance on construction, small errors in the diagram can lead to overall poor performance even if the citizen makes no interpretation errors. During (decision via) interpretation, background beliefs can again produce errors by causing the citizen to selectively reinterpret the meaning of the diagram syntax to produce conclusions consistent with her beliefs. The citizen may also simply make errors combining the different elements of the diagram to make causal inferences.

This study showed that the causal diagrams used in our task, which only represent the evidence provided in the text, do not capture all the knowledge citizens use to solve the problem. Because this is an ill-defined domain where we *want* citizens to make effective use of their background knowledge, we need to distinguish between normative and non-normative uses of background knowledge rather than asking citizens to check their common-sense at the door. Given that current tutoring systems ignore this problem either by prohibiting background knowledge or simply by not tutoring, some discussion of how we might address the problem is warranted.

It may be possible to provide automated tutoring on synthesis and to detect normative and non-normative uses of background knowledge by making three modifications: (a) providing a microworld to prevent speculation errors, (b) using confidence meters to detect and allow normative uses of background knowledge such as an override, and (c) using causal diagrams to represent both the evidence in the text as well as changes in the citizen's synthesized beliefs about the evidence so that the tutor can detect non-normative false-uncertainty, reverse-causation, and chaining errors. To illustrate how these modifications allow us to tutor deliberation, consider the following example.

During search, the deliberation tutor could provide a microworld that allows the citizen to conduct interviews or to collect experimental data from the sources in the text. This way, instead of allowing citizens to speculate about what a source might say, the tutor can require them to actually acquire that information.

Later, when the citizen is evaluating evidence the tutor can ask him to rate the strength of the evidence on a confidence meter–if he rates confirming studies more highly than disconfirming studies, then we can detect the error and provide feedback. Furthermore, if the citizen rates an anecdotal claim as stronger than an experiment, the tutor can enforce a basic constraint on evidence strength ratings, even without a fully specified normative theory of evidence evaluation. This approach allows us to prevent comprehension errors seen during diagram construction.

As the citizen starts to create a diagrammatic representation of the causal evidence, the tutor can provide traditional correctness feedback. Simultaneously, the tutor can again use a confidence meter to measure the citizen's synthesized belief about that causal relation. For example, perhaps the citizen begins the problem with 70% certainty that there is no relation between *junk food advertising* and *childhood obesity*, and then, after diagramming evidence showing an increase, incorrectly changes her synthesized belief by moving the confidence meter to 72% certainty that there is no relation. The tutor can provide feedback that she has changed her synthesized belief in the wrong direction. Likewise,

the tutor can also monitor bias in synthesis by ensuring that the citizen does not change her confidence more in response to confirming reports than to disconfirming reports. In this way, we partially allow the citizen to reason with her background knowledge, while still enforcing reasonable use of the evidence, allowing the tutor to correctly allow correct background knowledge to override weaker evidence.

At this point, the tutor has ensured that the citizen's synthesized beliefs reflect a reasonable synthesis of his background knowledge with the evidence provided. When the citizen must finally interpret the diagram to make a policy decision, he does so using a single synthesized model from which the tutor can detect non-normative interpretation errors (i.e., chaining, reverse causation, and false uncertainty errors).

In this manner, the deliberation tutor can provide feedback while allowing citizens who start with two different sets of background knowledge to construct two different, but reasonable, synthesized models, and reason to two different solutions. Thus by designing a tutor to address the errors found in the second study, we may tutor deliberation tasks that have the key characteristics of ill-defined problems: large search spaces, multiple representations, and multiple correct answers–the holy grail of tutoring in ill-defined domains.

CONCLUSION

The purpose of this project was to determine: (a) whether causal diagrams improve deliberation compared to text, (b) whether construction promotes learning, and (c) the learning difficulties associated with constructing and interpreting causal diagrams.

With respect to the first question, we found that having a correct causal diagram improves deliberation, supporting the conjecture that causal diagrams can improve deliberation and thus provide a good representation system for a deliberation tutor. This advances our immediate project to build a deliberation tutor and more generally addresses the lack of research on causal diagrams.

With respect to the second question, we found that students who had practiced constructing causal diagrams were better prepared for future deliberation than students given diagrams or text, even though these students did not later construct diagrams. This result is surprising considering that students received virtually no instruction or feedback on constructing diagrams, and considering the research showing no benefit of construction. It is possible that practice constructing diagrams improves comprehension skills that can be used later even when one returns to a text-based strategy. Most studies comparing diagrams and text would not observe this result if they do not test for the effects of construction practice on transfer. This result shows that there are differential effects of receiving and construction diagrams on performance and transfer.

With respect to the third question, we found that the learning difficulties that pose the greatest challenge for a deliberation tutor are the normative uses of background knowledge during diagram interpretation, which must be allowed but which must be distinguished from non-normative uses of background knowledge and simple errors. A deliberation tutor might overcome this challenge by monitoring both the student's representation of the evidence and the student's representation of his synthesized beliefs about the evidence.

To conclude, this work contributes to deliberation tutoring by identifying causal diagrams as an effective representation system, by showing that practice constructing diagrams improves future deliberation and by identifying the nuanced ways in which a tutor must monitor background

knowledge. These findings apply not only to deliberation, but to all domains that rely on causal reasoning including natural science, history, strategic planning, and medicine and more generally to domains in which people must construct representations of conflicting evidence from multiple sources such as argument, law, and intelligence.

ACKNOWLEDGEMENTS

The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B040063 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

REFERENCES

Ainsworth, S. E. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction,* 16(3), 183-98.

Britt, M. A., & Aglinskas, C. (2002). Improving students' ability to identify and use source information. *Cognition and Instruction,* 20(4), 485-522.

Britt, M. A., Rouet, J. F., Georgi, M. C., & Perfetti, C. A. (1994). Learning from history texts: From causal analysis to argument models. In G. Leinhardt, I. L. Beck, & C. Stainton (Eds.) *Teaching and Learning in History* (pp. 47-84). Hillsdale, NJ: Lawrence Erlbaum.

Carnegie Corporation of New York, & CIRCLE: The Center for Information & Research on Civic Learning & Engagement (2003). *The civic mission of schools.* New York: Carnegie Corp. of New York.

Carr, C. S. (2003). Using computer supported argument visualization to teach legal argumentation. In P. A. Kirschner, S. J. Buckingham Shum, & C. S. Carr (Eds.) *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-making* (pp. 75-96). London: Springer-Verlag.

Chinn, C. A., & Brewer, W. F. (2001). Models of data: A theory of how people evaluate data. *Cognition and Instruction,* 19(3), 323-93.

Cox, R. (1996). *Analytical reasoning with multiple external representations.* Unpublished doctoral dissertation, University of Edinburgh, England.

Easterday, M. W., Aleven, V., & Scheines, R. (2007a). 'Tis better to construct or to receive? Effect of diagrams on analysis of social policy. In R. Luckin, K. R. Koedinger, & J. Greer (Eds.) *Proceedings of the 13th International Conference on Artificial Intelligence in Education* (pp. 93-100). Amsterdam: IOS Press.

Easterday, M. W., Aleven, V., & Scheines, R. (2007b). The logic of Babel: Causal reasoning from conflicting sources. In V. Aleven, K. Ashley, C. Lynch, & N. Pinkwart (Eds.) *Proceedings of the Workshop on AIED Applications of Ill-Defined Domains at the 13th International Conference on Artificial Intelligence in Education* (pp. 31-40). Los Angeles, CA.

Easterday, M. W., Alevan, V., Scheines, R., & Carver, S. M. (2009). Will Google destroy western democracy? Bias in policy problem solving. In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A. Graesser (Eds.) *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (pp. 249-256). Amsterdam: IOS Press.

Easterday, M. W., Kanarek, J., & Harrell, M. (2009). Design requirements of argument mapping software for teaching deliberation. In T. Davies, & S. P. Gangadharan (Eds.) *Online Deliberation: Design, Research, and Practice.* Stanford, CA: Center for the Study of Language and Information.

Fishkin, J. S. (1995). *The voice of the people: Public opinion and democracy.* London: Yale University Press.

Forbus, K. D., Carney, K., Sherin, B. L., & Ureel, L. C. (2005). VModel: A visual qualitative modeling environment for middle-school students. *AI Magazine,* 26(3), 63-72.

Graesser, A. C., Wiley, J., Goldman, S. R., O'Reilly, T., Jeon, M., & McDaniel, B. (2007). SEEK web tutor: Fostering a critical stance while exploring the causes of volcanic eruption. *Metacognition Learning,* 2, 89-105.

Grossen, B., & Carnine, D. (1990). Diagramming a logic strategy: Effects on difficult problem types and transfer. *Learning Disability Quarterly,* 13(3), 168-82.

Hall, V. C., Bailey, J., & Tillman, C. (1997). Can student-generated illustrations be worth ten thousand words? *Journal of Educational Psychology,* 89(4), 677-81.

Harrell, M. (2008). No computer program required: Even pencil-and-paper argument mapping improves critical thinking skills. *Teaching Philosophy, 31, 351-374.*

Heuer, R. J. (1999). *Psychology of intelligence analysis.* New York: Novinka Books.

Horn, R. E., & Weber, R. P. (2007). New tools for resolving wicked problems: Mess mapping and resolution mapping processes. Retrieved from http://www.strategykinetics.com/files/New_Tools_For_Resolving_Wicked_Problems.pdf

Huff, A. S., & Jenkins, M. (Eds.). (2002). *Mapping Strategic Knowledge*. London: Sage.

Joffe, M., & Mindell, J. (2006). Complex causal process diagrams for analyzing the health impacts of policy interventions. *American Journal of Public Health,* 96(3), 473-9.

Jonassen, D., & Ionas, I. (2008). Designing effective supports for causal reasoning. *Educational Technology Research and Development,* 56(3), 287-308.

Jones, D. K., & Read, S. J. (2005). Expert-novice difference in the understanding and explanation of complex political conflicts. *Discourse Processes,* 39(1), 45-80.

Kirschner, P. A., Buckingham Shum, S. J., & Carr, C. S. (Eds.). (2003). *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-making*. London: Springer-Verlag.

Koedinger, K. R., & Nathan, M. J. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *The Journal of the Learning Sciences,* 13(2), 129-64.

Kuhn, D. (2005). *Education for Thinking.* Cambridge, MA: Harvard University Press.

Kuhn, D., & Dean Jr, D. (2004). Connecting scientific reasoning and causal inference. *Journal of Cognition and Development,* 5(2), 261-88.

Kuhn, D., Amsel, E., & O'Loughlin, M. (1988). *The Development of Scientific Thinking Skills.* San Diego: Academic Press.

Kuipers, B., & Kassirer, J. P. (1984). Casual reasoning in medicine: Analysis of a protocol. *Cognitive Science,* 8(4), 363-85.

Leelawong, K., & Biswas, G. (2008). Designing learning by teaching agents: The Betty's Brain system. *International Journal of Artificial Intelligence in Education,* 18(3), 181-208.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology,* 37(11), 2098-109.

Lynch, C. F., Ashley, K. D., Aleven, V. A., & Pinkwart (2006). Defining "ill-defined domains": A literature survey. In V. Aleven, K. Ashley, C. Lynch, & N. Pinkwart (Eds.) *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains at the 8th International Conference on Intelligent Tutoring Systems* (pp. 1-10). Jhongli (Taiwan), National Central University.

Masterman, L. (2005). A knowledge-based coach for reasoning about historical causation. In C. K. Looi, G. McCalla, B. Bredeweg, & J. Breuker (Eds.) *Proceedings of the 12th International Conference on Artificial Intelligence in Education* (pp. 435-42). Amsterdam: IOS Press.

Mayer, R. E. (2001). *Multimedia learning.* Cambridge, United Kingdom: Cambridge University Press.

McCrudden, M. T., Schraw, G., Lehman, S., & Poliquin, A. (2007). The effect of causal diagrams on text learning. *Contemporary Educational Psychology,* 32(3), 367-88.

Montibeller, G., & Belton, V. (2006). Causal maps and the evaluation of decision options-a review. *Journal of the Operational Research Society,* 57(7), 779-91.

Narayanan V. K., & Armstrong D. J. (Eds.). (2005). *Causal Mapping for Research in Information Technology*. Hershey, PA: Idea Group.

Patton, C. V., & Sawicki, D. S. (1993). *Basic Methods of Policy Analysis and Planning.* Upper Saddle River, NJ: Prentice Hall.

Pawson, R. (2006). *Evidence-based Policy: A Realist Perspective.* London: Sage.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference.* Cambridge, UK: Cambridge University Press.

Perfetti, C. A., Rouet, J. F., & Britt, M. A. (1999). Toward a theory of documents representation. In H. van Oostendorp, & S. R. Goldman (Eds.) *The Construction of Mental Representations During Reading* (pp. 99-122). Mahway, NJ: Lawrence Erlbaum.

Pinkwart, N., Aleven, V., Ashley, K., & Lynch, C. (2007). Evaluating legal argument instruction with graphical representations using LARGO. In R. Luckin, K. R. Koedinger, & J. Greer (Eds.) *Proceedings of the 13th International Conference on Artificial Intelligence in Education* (pp. 101-8). Amsterdam: IOS Press.

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences,* 4, 155-69.

Salles, P., Bredeweg, B., & Araújo, S. (2006). Qualitative models about stream ecosystem recovery: Exploratory studies. *Ecological Modelling,* 194(1-3), 80-9.

Scheines, R., Easterday, M., & Danks, D. (2007). Teaching the normative theory of causal reasoning. In A. Gopnik, & L. Schultz (Eds.) *Causal learning: Psychology, Philosophy, and Computation* (pp. 119-38). Oxford, England: Oxford University Press.

Simon, H. A. (1996). Social planning: Designing the evolving artifact. In H. A. Simon (Ed.), *The Sciences of the Artificial (3rd ed.)* Cambridge, MA: MIT Press.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, Prediction, and Search (2nd ed.).* Cambridge, MA: MIT Press.

Stull, A. T., & Mayer, R. E. (2007). Learning by doing versus learning by viewing: Three experimental comparisons of learner-generated versus author-provided graphic organizers. *Journal of Educational Psychology,* 90(4), 808-20.

Suthers, D. D., & Hundhausen, C. D. (2003). An experimental study of the effects of representational guidance on collaborative learning processes. *The Journal of the Learning Sciences,* 12(2), 183-218.

Suthers, D. D., Weiner, A., Connelly, J., & Paolucci, M. (1995). Belvedere: Engaging students in critical discussion of science and public policy issues. In J. Greer (Ed.), *Proceedings of the 7th World Conference on Artificial Intelligence in Education* (pp. 266-273). Charlottesville, VA: Association for the Advancement of Computing in Education.

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science,* 50(3), 755-69.

Twardy, C. R. (2004). Argument maps improve critical thinking. *Teaching Philosophy,* 27(2), 95-116.

Van Gelder, T. J. (2003). Enhancing deliberation through computer supported visualization. In P. A. Kirschner, S. J. Buckingham Shum, & C. S. Carr (Eds.) *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-making* (pp. 97-115). London: Springer-Verlag.

Voss, J. F. (2005). Toulmin's Model and the solving of ill-structured problems. *Argumentation,* 19(3), 321-9.

Voss, J. F., Carretero, M., Kennet, J., & Silfies, L. N. (1994). The collapse of the soviet union: A case study in causal reasoning. In M. Carretero, & J. F. Voss (Eds.), *Cognitive and Instructional Processes in History and the Social Sciences* (pp. 403-29). Hillsdale, New Jersey: Lawrence Erlbaum.

Voss, J. F., Tyler, S. W., & Yengo, L. A. (1983). Individual differences in the solving of social science problems. In R. F. Dillion, & R. R. Schmeck (Eds.) *Individual Differences in Cognition* (pp. 205-32). New York: Academic Press.

Walker, W, & Fisher, G. (1994). *Public Policy Analysis: A Brief Definition (Document. No. P-7856).* Santa Monica, CA: RAND Corporation.

Zimmerman (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review,* 27(2), 172-223.