A Cognitive Game for Teaching Policy Argument

Matthew W. Easterday

August 2010
CMU-HCII-10-106

Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Richard Scheines (Co-chair)
Vincent Aleven (Co-chair)
Sharon Carver (Co-chair)
Gautam Biswas (Vanderbilt University)

Submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy and the Program for Interdisciplinary Educational Research

Abstract

Our democracy depends upon the creation of an active engaged citizenry. The purpose of this dissertation is to provide the foundational research necessary for constructing an intelligent tutoring system to teach policy deliberation. The dissertation makes five use-inspired basic research contributions to the knowledge and technology of Intelligent Tutoring Systems and Artificial Intelligence in Education. Specifically it: (a) develops a cognitive framework for deliberation, (b) localizes reasoning difficulties within the synthesis stage of the framework, (c) shows that causal diagrams can improve reasoning, (d) demonstrates that we can design intelligent tutoring systems that teach deliberation, and (e) shows that educational games can increase learning and interest by using intelligent tutoring approaches to providing assistance.

Acknowledgements

This work was supported in part by a graduate training grant awarded to Carnegie Mellon University by the U.S. Department of Education (# R305B040063), the Siebel Scholars Foundation, and the Pittsburgh Science of Learning Center, which is funded by the National Science Foundation (# SBE-0836012). The opinions expressed are those of the author and do not represent the views of the U.S. Department of Education, the Siebel Scholars Foundation, or the National Science Foundation.

Table of contents

1. Democratizing deliberation through intelligent tutoring

Summary. Our democracy depends on the creation of an active engaged citizenry. The purpose of this dissertation is to provide the foundational research necessary for constructing and an intelligent tutoring system for policy deliberation. This chapter describes five use-inspired basic research contributions to the knowledge and technology of Intelligent Tutoring Systems and Artificial Intelligence in Education. These contributions are: (a) to develop a cognitive framework for deliberation, (b) to localize reasoning difficulties within the framework, (c) to show that causal diagrams can improve reasoning, (d) to demonstrate that deliberation can be tutored, and (e) to show that games can better increase learning and interest by using intelligent tutoring approaches to providing assistance.

We are continually beset by political problems. As this dissertation is being written, we are facing the worst economic crisis since the Great Depression, we are experiencing the worst oil spill in the nation's history as the threat of global warming continues unabated; and, we are fighting two wars in Iraq and Afghanistan the latter of which is now the longest war ever waged by the United States. Our political system seems unable to prevent these catastrophes or to act decisively after crises emerge.

Underlying these problems is a weakened democracy. If we were to try to define the "operating system" of our democracy: its ability to detect problems (journalism), its ability to make decisions about problems (an active engaged citizenry), and its ability to act on these decisions (representatives uncorrupted by financial influence), we can see that it does not function as well as one might hope. The institution of commercial journalism is slowly imploding, only about half the population is willing to vote (the most minimal form of political participation), and our elected officials spend the bulk of their time fundraising with a clear effect on their behavior. But we cannot even hope to repair these elements if we lack a basic foundation of civic education–a civic education that teaches students how to become active engaged citizens who can deliberate, persuade, and act.

The research goal of this dissertation is **to provide the foundational research necessary to construct an intelligent tutoring system for democratizing civic knowledge;** that is, to help citizens learn to deliberate using software that provides automated assistance to learners (VanLehn, 2006; Woolf, 2009). This research and development question holds relevance to several fields of knowledge yet remains virtually unaddressed.

Learning Science

Psychologists have for the most part ignored one of the most pressing educational imperatives of our time. This is despite the fact that creating effective instruction in policy deliberation relies heavily on issues central to cognitive science including: confirmation bias, external representations, feedback, and problem solving in ill-defined domains. Having shown that citizens do not deliberate rationally on the basis of evidence, some psychologists have even argued that leaders should appeal more to emotion or better frame the issues (Westen, 2007; Lakoff, 2002). While more persuasion might win elections, it will do little to promote an active, engaged citizenry that demands better policy. The alternative, ignored by cognitive science, is civic education (The Center for Information and Research on Civic Learning and Engagement & The Carnegie Corporation of New York, 2003).

It is not surprising that psychologists have ignored this research problem. The development of a deliberation tutor primarily concerns the *design* of *instructional dynamics* that is, interactions between students, teachers, content, and environment. Even schools of education for which the design of instructional dynamics should be the primary focus, frequently ignore these issues (Ball & Forzani, 2007). In addition, the problem of deliberation tutoring requires resources in artificial intelligence and software engineering that are not always available to schools of education.

Sciences of the Artificial, HCI, and Intelligent Tutoring research

The question of developing an intelligent deliberation tutor also falls squarely within the purview of Human-Computer Interaction (HCI). HCI has been defined as the design and study of the artificial, i.e., man-made artifacts (Simon, 1996). HCI can also be thought of as the set of problems that involve the interaction of audience with technology to achieve a purpose (Buchanan, 1999). From Simon's perspective, we see that research on intelligent tutoring systems (ITS) clearly concerns the design and study of a particular form of the artificial. From Buchanan's perspective, we see that ITS concerns the interaction between an audience (learners), and a technology (intelligent tutoring systems) to achieve a purpose (expertise). In fact, the focus of this particular research question on policy and intelligent tutoring is one of the central themes in Simon's *Sciences of the Artificial*, which devotes a great deal of time to the design problems of social planning. This research question also falls under one of the grand challenges of information science: to provide a teacher for every student (Computing Research Association, 2003; 2005).

Sadly, the intelligent tutoring community has produced little work on deliberation tutoring. This is most likely for two reasons. First, while deficits in civic skill may be the undoing of our republic, the importance of deliberation is not reflected in research spending, which emphasizes reading and STEM (science, technology, engineering and mathematics). Second, the ill-structured nature of deliberation has made the development of a viable deliberative tutoring system difficult given our current capabilities in artificial intelligence. So while the ITS community is well poised to address this research question, practical considerations push ITS research in other directions.

Pasteur's Quadrant: Use-inspired, basic research

The Post-WWII, U.S. model of government-sponsored research advocated by Vannever Bush holds that our nation's health, prosperity, and security depend on the technological advances of applied research, which in turn depends on basic research (Bush, 1945). This model led to the creation of the National Science Foundation.

However, other studies examining the design of technological systems find that basic research may play a limited role in the development of a specific technology (Sherwin & Isenson, 1967). The more modern Pasteur's Quadrant model of research argues that for the nation to capture the technological return on its investment in basic science, we must embrace a third type of research: use-inspired, basic-research, "… work that locates the center of research in an area of basic scientific ignorance that lies at the heart of a social problem" (Stokes 1997). There are an increasingly large number of examples of Pasteur's Quadrant research. Edwin Land, inventor of the Polaroid, described his approach in this manner:

> If you sense a deep human need, then you go back to all the basic science. If there is some missing, then you try to do more basic science and applied science until you get it. So you make the system to fulfill that need … (McElheny, 2002, p. 115.)

This problem of how to create an intelligent tutor for deliberation, is unabashedly use-inspired and primarily a problem of *design* rather than understanding. The long term purpose of this research is to design a piece of technology that meets a social need.

Many researchers, especially those in psychology, and even some from the ITS community may dismiss this as merely an "engineering" or "practical" problem, i.e., not research at all. But this demonstrates a certain misconception about the nature of the sciences of the artificial, including HCI and educational research. If we take the Pasteur's Quadrant approach seriously, then we see that in some cases, the scientific questions of basic research may only be a means to an end. Contributions to basic research and scientific understanding are produced in the process of developing new technology, but the purpose, the *research* goal, is to develop a system that meets a human need. Land explains the basic logic:

> You always start with a fantasy. Part of the fantasy technique is to visualize something as perfect. Then with the experiments you work back from the fantasy to reality, hacking away at the components. (McElheny, 2002, p. 115.)

The design problem is research, because the scientific knowledge required to construct the given system is unknown.

If educational research is to address questions of use, then research that only addresses questions of understanding will not be sufficient. A fully specified learning theory, consisting of a set of empirically-based design principles, would still offer no clear path to effective instructional systems. Research on instructional systems is also necessary, because it provides worked-examples showing how specific designs can overcome all the relevant design challenges and allows for future incremental improvement. In fact, these worked examples may prove more useful than even a fully described theory of the learning principles. To illustrate this point: imagine that you are trying to design an airplane, and you've never seen one before – which would be more useful: another working airplane, or a fully specified theory of aerodynamics? Furthermore, even if one is concerned only with principles, it's hard to see how these principles could be tested without first building working systems (this is why learning scientists such as Nathan and Alibali (2010) advocate *scaling down*, as opposed to scaling up). Both systems and principles are necessary, but working systems have perhaps been undervalued given our intellectual roots in cognitive science.

Contributions

The contribution of this dissertation is to demonstrate the viability of an intelligent tutoring system for teaching deliberation. While the research agenda is driven by problems of use and design, this work makes several contributions in basic understanding, as well as to the fields of education, human-computer interaction, and intelligent tutoring systems. In the process of "hacking away at the components" of this use-inspired research problem, this dissertation will:

1. **Develop a cognitive framework for deliberation**. Previous research on deliberation proposed models of argumentation that focused on how people reason from recall. These models did not

attempt to address: how people should incorporate new information, how people make decisions based on evidence, or how to provide automated tutoring. The cognitive framework for deliberation proposed in Chapter 2 provides a grammar for thinking and communicating about policy reasoning. The empirical propositions in this work can all be defined in terms of the framework, e.g., where students have difficulties, where the theoretical gaps in previous work lie, and what future goals to focus on. With respect to practice, the cognitive framework for deliberation defines standards for what civic educators should teach. Each chapter of the dissertation will demonstrate how the framework can be used to concisely describe questions and results and more importantly, how to guide design.

2. **Localize reasoning difficulties in synthesis.** Previous research predicts that we should see bias in how students *search* for and *evaluate* evidence. However, for the policy tasks used here, the biggest learning challenge, the relatively minor impact evidence has on prior beliefs, seems to occur in the *synthesis* of evidence (Chapter 3). This suggests that research and instruction in deliberation should focus on teaching *synthesis*.

3. **Show that causal diagrams can improve policy reasoning.** Previous research makes no strong predictions about whether we can effectively use external representations to improve policy reasoning. It also provides little guidance in how to design an effective representation. Research also predicts that constructing diagrams is at best a necessary evil (in the case where a diagram is required but not provided) which does not promote learning. However, Chapter 4 shows that not only does providing students with a correct causal diagram improve policy reasoning, practicing constructing diagrams helps improve future reasoning even when diagrams are unavailable. Chapter 4 also demonstrates some of the serious challenges in learning to construct diagrams. These results suggest that causal diagrams should be used for deliberation, and that we must devote significant effort to understanding how to teach diagram construction.

4. **Demonstrate that deliberation can be tutored.** Some have argued that policy problems are *wicked,* i.e., undefinable, and the intelligent tutoring systems community has had limited success tutoring argumentation or causal reasoning. Probably the single most important contribution of this work is that it shows how to design an instructional system for policy reasoning that provides a level of cognitive feedback approaching that of cognitive tutors for algebra (Chapter 5). This is made possible because the cognitive model breaks the task into semi-structured steps, the use of external representations forces students to represent knowledge in a machine-readable form, and the use of causal diagrams allows the tutor to make inferences about the students' beliefs that are not possible using the more popular Toulmin/Beardsley argument diagrams. Theoretically, this opens up tutoring in a variety of ill-defined/media literacy/computational literacy domains where some ill-formed set of information must be transformed into a formal model and used to improve reasoning. Such domains include argumentation, law, history, contextual modeling, and lesson study. Practically, this instructional system shows that we can provide automated tutoring of deliberation, a major step toward the democratization of civic skill.

5. **Show that games can better increase learning and motivation with tutor-like assistance.** There are almost no randomized controlled experiments comparing the effectiveness of tutors and games. Chapter 6 shows that not only can we combine a game environment with a tutoring system, but that using a more tutor-like assistance increases both learning and interest. This suggests that educational game designers should consider using intelligent tutoring systems. It

also suggests that intelligent tutoring developers can reap the motivational benefits of fantasy environments without substantially altering how assistance is provided.

The research agenda for the dissertation is as follows: I will define a cognitive model for deliberation (Chapter 2), identify bias within the deliberation task (Chapter 3), test how diagrams might overcome bias and the problem of multiple representations, and investigate the difficulties associated with learning to use diagrams (Chapter 4), design an intelligent tutor that can provide assistance in learning deliberation (Chapter 5), and test the relative effectiveness of combining tutors and games (Chapter 6). The dissertation will conclude with a discussion of how this work can then be carried forward in order to create a full curriculum for engaged citizenship.

2. Research on learning environments for deliberation

Summary. What does prior research tell us about building a tutor for deliberation? We can decompose the problem by defining a *learning environment platform* whose layers specify the knowledge and technology needed to create an evidence-based learning environment. These layers include: (a) the *task* we want to teach, (b) a *cognitive model* describing the expertise needed to perform those tasks, (c) the *learning elements* including the *learning challenges* a novice faces in acquiring expertise, and the *learning principles* applied to overcome these challenges, (d) an *instructional system* for teaching that includes a *delivery system*, an *inquiry environment,* and *assistance,* and (e) a *curriculum* that combines the instructional systems into a coherent environment. A review of the literature relevant to each layer of a learning environment platform that teaches policy deliberation leads to the five research questions asked in this dissertation. These questions include: (a) what cognitive model best describes deliberative skill?, (b) what learning challenges do students face in acquiring this cognitive model?, (c) can causal diagrams help students overcome difficulties in policy reasoning?, (d) is it better to provide game-like or tutor-like assistance?, and (e) how can we develop an instructional system that combines a game-like inquiry environment with an intelligent tutoring system to teach deliberation? To answer the first question, this chapter proposes a task-analytical, cognitive model of deliberation. This model proposes that solving simple deliberation problems requires several sets of skills including: (a) defining a question, (b) searching for information, (c) evaluating evidence, (d) constructing an external representation of the problem, (e) synthesizing the evidence, and (f) interpreting the external representation to make a policy decision.

How can we design a tutoring system that can teach the skills of deliberation? What "components must be hacked away" to create such a system? Unlike educational research in math, reading, and science, there is no established research effort to build upon, and no previous instructional systems from which incremental improvements can be made. It is not even clear what components are required to make this vision a reality.

If research from other domains does not tell us how to construct the components of a deliberation tutor, then at least we can use it to make guesses about which components are needed. We *can* make use of previous research from other domains to define what a solution might look like in the abstract. The goal of this chapter is to determine which components can be constructed using our current store of scientific knowledge, and which must be hacked away. These components together will form a *learning environment platform*.

In the abstract, a ***learning environment platform*** (Figure 2.1) for policy reasoning consists of 5 components or layers: (i) the reasoning tasks an expert policy reasoner should be able to perform, (ii) the cognitive models defining the abilities of the expert and student, (iii) the learning elements which include both the learning challenges faced by a particular group of students, and the instructional principles that might be applied to overcome these challenges, (iv) instructional systems for teaching policy reasoning (in this case software) which include: a delivery system, an inquiry environment for problem solving, and a means of providing assistance, and (v) a curriculum that combines the instructional systems into a coherent learning environment.

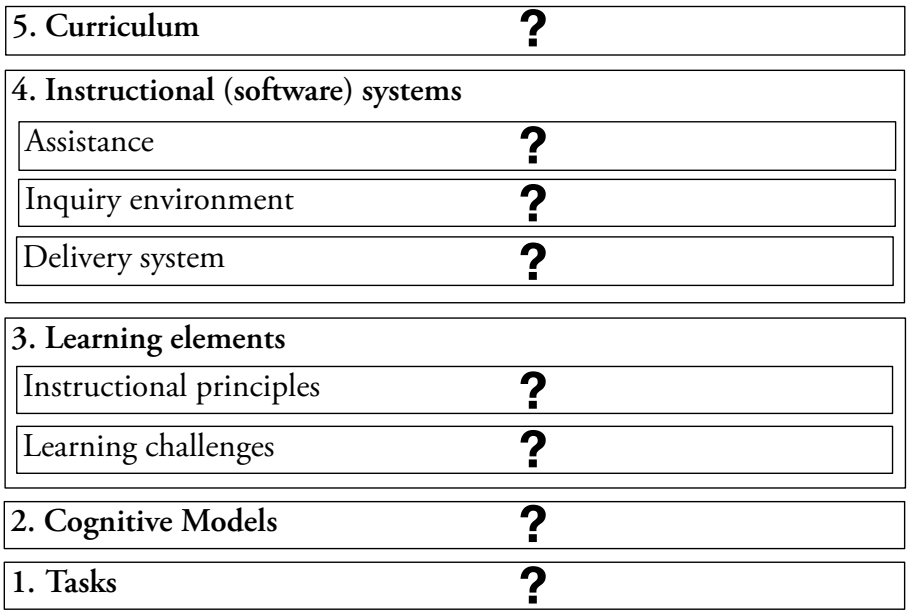


Figure 2.1. Layers of a learning environment platform.

What does previous research tell us about each of the layers when it comes to policy reasoning? In hacking away at these layers, five research questions arise:

1. What cognitive model best describes deliberative skill?
2. What learning challenges do students face in acquiring this cognitive model?
3. Can causal diagrams help students overcome difficulties in policy reasoning?
4. (If we use a combination of games and tutors to teach deliberation), is it better to provide game-like or tutor-like assistance?
5. How can we develop an instructional system that combines a game-like inquiry environment with an intelligent tutoring system to teach deliberation?

I first consider layer 1: the policy reasoning task, and what current research tells us about it.

Layer 1 :: Tasks : A policy reasoning task

What is a policy reasoning task? In the few psychological studies of policy reasoning, participants are typically asked to explain their synthesized model of the policy problem or to use their model to justify a solution. For example Kuhn (1991) asked people to explain what causes kids to fail in school, Voss, Tyler & Yengo (1983) asked people how they would improve Soviet agricultural productivity if they were the USSR Minister of Agriculture, Jones & Read (2005) asked experts and novices to talk and answer questions about the Israel-Palestine conflict, and Axelrod (1976) analyzed verbal protocols of politicians. A second type of study presents people with new information in order to see how their model changes. For example, Lord, Ross, & Lepper (1979) present students with evidence about the result of a study on the death penalty, Kuhn (1991) presented people with several types of confounded or non-evidence, and Taber & Lodge (2006) allowed students to search

for arguments about topics like affirmative action from a list of known sources like the Republican Party.

We can combine these studies to provide us with a simple but complete policy reasoning task. The reasoner is provided with: an initial policy question like: *what should we do about global warming?*, access to a set of information about that question, and then must propose a solution to the problem with some justification for that solution. We could of course imagine more sophisticated versions of this task, for example we could allow the reasoner to conduct empirical studies to add to their set of information resources, we could require that the solution satisfies the goals of multiple conflicting parties, or we could enforce temporal constraints on problem solving. However, the simple version of the task is both consistent (and of far larger scope!) than previous cognitive studies, as well as consistent with normative descriptions of policy analysis (e.g. Pawson 2006). So we can use this simplified task as the starting point.

With this policy reasoning task in hand, how does previous research say one one should go about solving it?

Layer 2 :: Cognitive model: A framework for deliberation

The learning environment platform rests on a cognitive model of problem solving. Defining this layer raises the first research question considered in this dissertation: *what cognitive model best describes deliberative skill*?

A simple task analytic framework for deliberation is shown in Figure 2.2, consisting of several processes including: questioning, searching for information, comprehending and evaluating evidence, synthesizing evidence, and deciding. 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.

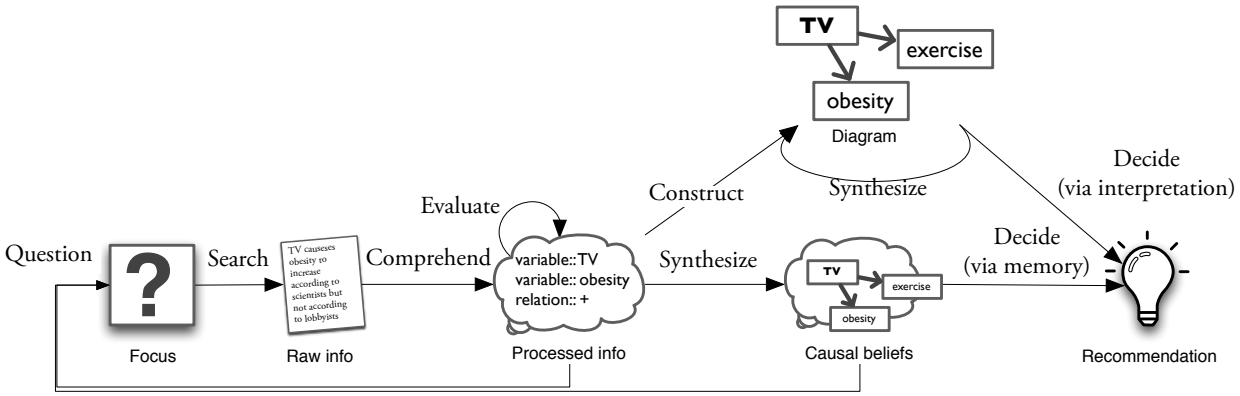


Figure 2.2. A cognitive framework for deliberation. The framework roughly defines the stages of problem solving with and without diagrams.

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, 2007; 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 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 consensus on a normative theory for evaluating evidence, and there certainly is no normative theory that provides intersubjective criteria for quantitatively weighing evidence from different studies, i.e., how much an observational study is *worth* compared to an experiment or to a mechanistic explanation. Impartial evidence evaluation of policy information proves difficult (Taber & Lodge, 2006, Lord, Ross, & Lepper 1979, Kuhn et al. 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., 2007; 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 health care costs), 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. 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., 2007; Voss, 2005; Horn & Webber, 2007; Rittle & Webber, 1973). Note that this does *not* mean that *any* answer is as good as another.

The deliberation framework delineates 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. In doing so, it 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 educational technology research across domains.

This analytical task analysis provides the first contribution of the dissertation: an initial cognitive model of deliberation.

Layer 3 :: Learning elements : Bias, diagrams, feedback, and games

At this point, I have defined the policy task and a model of deliberation outlining how the task should be solved. The next layer of the learning environment platform includes the learning elements, which include: (a) learning challenges, and (b) instructional principles. Where does previous research predict that students will encounter difficulty acquiring this model, and what instructional principles can be used to overcome these difficulties? Given the lack of research on learning deliberation, we must speculate from work on other domains.

One learning challenge we should certainly expect is confirmation bias.

Learning challenge: Bias

Previous research predicts that bias will present a significant learning challenge, but we do not know exactly where, since we know little about policy reasoning in general. Indeed, bias presents such a problem that some psychologists, having shown that citizens do not reason rationally on the basis of evidence about policy issues such as global warming or financial regulation, have argued that leaders should appeal more to emotion or better frame the issues (Westen, 2007; Lakoff, 2002). While

more effective persuasion might win elections, it will do little to promote an active, engaged citizenry that demands better policy. The alternative of course is civic education (CIRCLE, 2003). As a preliminary step toward a cognitive tutor for civics, the study presented here examines where bias occurs during a policy reasoning task (Kuhn 1991; Voss, Greene, Post & Penner 1983; Voss, Tyler & Yengo 1983).

A policy reasoning task, like deciding whether decreasing classroom size will increase school performance, requires reasoners to decide whether a policy will lead to the desired outcome on the basis of evidence. Unfortunately, psychologists have shown a consistent pattern of confirmation bias, where the reasoner's prior beliefs overwhelm impartial evaluation of evidence (Nickerson, 1998; MacCoun, 1998; Kunda 1990). On emotional topics such as the death penalty and gun control, psychologists have even shown that partisans from opposite sides can each become more convinced of their original position after seeing the same set of evidence (Lord, Ross, & Lepper, 1979). Bias arises both during search (Taber, & Lodge, 2006; Redlawsk, 2002) and during analysis of information (Kuhn, Amsel, & O'Loughlin, 1988; Koslowski, 1996; Chinn & Brewer, 2001; Zimmerman, 2000; MacCoun, 1998).

Most of the studies on policy reasoning have used tasks in which students evaluate arguments (Kuhn, 1991; Taber & Lodge, 2006) as opposed to empirical evidence (Lord, Ross, & Lepper, 1979) focusing on analysis (Kuhn, 1991) rather than search (Taber & Lodge, 2006; Redlawsk, 2002).

This leads us to the second research question: *what learning challenges do students face in acquiring the cognitive model of deliberative skill?*

Instructional Principle: Causal diagrams

What instructional tactics might we use to overcome the challenges of complexity and bias? There are a number we might look at. The first is to use external representations.

The cognitive 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 2.3 (left).

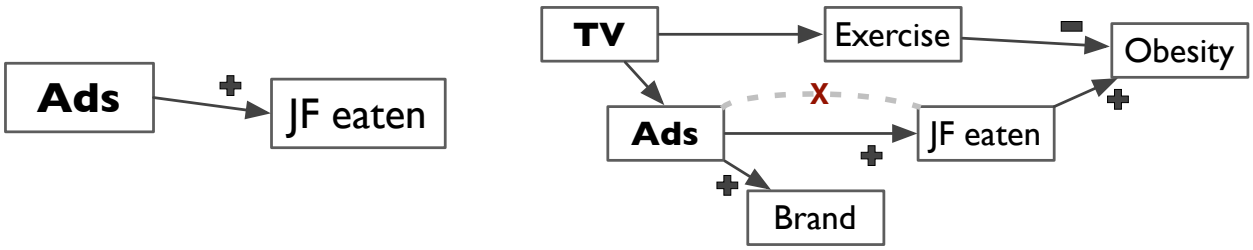


Figure 2.3. A diagram element representing the causal claim that advertising increases junk food eaten (left) and a whole causal model of advertising and obesity (right).

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 2.3 (right).

In the early phases of problem solving, the citizen created a diagrammatic representation of the problem that she must now **interpret** by balancing values, weighing the costs and benefits of different interventions, and deciding upon the "best" intervention. These tasks correspond to inferences made from the the causal diagram, i.e., identify 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).

Diagrams might help to overcome some policy reasoning difficulties in the same way that equations help us solve algebra problems. External representations can relieve the memory burden required when using multiple pieces of information to solve a problem. In the case of diagrams, they may also allow us use our visual processing abilities to make cognitive inferences, i.e., to *see* an answer to problem by looking at the diagram. The benefits of diagrams do not come "for free" however. Learning to use a certain type of diagram can impose its own set of learning challenges. Diagram users must both learn how to construct the representation and how to read its symbols. As we investigate the effectiveness of diagrams as an instructional tactic at the *learning principles* level, we will also move back down to the *learning challenges* level to identify barriers students face in learning to use diagrams.

For causal diagrams to improve performance, 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 construct and interpret the diagram. 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, Aleven, Ashley, & Lynch 2007; Harrell, in press; Twardy, 2004), no different (Carr, 2003) or even more difficult (Koedinger & Nathan, 2003) 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).

Diagrams might be useful for increasing learning if they help the student to develop knowledge and skills during training that can be used later even when diagrams are unavailable. For example, perhaps the process of constructing diagrams helps students learn to encode causal claims whether or not they use the diagram later. 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 may have considerable difficulty constructing diagrams (Cox, 1996) and learning to construct diagrams may require extensive training (Grossen & Carnine, 1990). By an analogous 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 understanding of the material being diagrammed, 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 construct the diagram, and how to make inferences from 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 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, Shum, & Carr, 2003), intelligence analysis (Heuer, 1999), and legal reasoning (Pinkwart, Aleven, Ashley, & Lynch, 2007).

This raises a second issue. 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 (Kirischner 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, I 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.

Causal diagrams seem like a good bet, but it's clear that during synthesis we confront the problem of multiple representations, i.e., that there is no agreed upon representation of this 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. (2007) 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 2.3, 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.

Testing which representation systems best improve deliberation, we may eventually reach consensus on how to teach deliberation. This issue leads to the third research question: *can causal diagrams help students overcome difficulties in policy reasoning?*

Instructional principle: Combining tutors and games

A second instructional principle we might employ is to combine intelligent tutors with games. Intuitively we believe that tutors are good for learning and games are good for fun, so perhaps there is a way to combine them and get the best of both.

Unfortunately, our approaches to designing educational games and our principles for designing intelligent tutors each lead us to a different set of conflicting instructional systems. Some of these design conflicts can be easily resolved, others are more difficult. For example, whereas video games typically use fantasy environments, tutors usually do not. We know that adding fantasy contexts to educational games can improve both learning and interest (Cordova & Lepper, 1996), so embedding a tutor in a game-like fantasy context seems like a straightforward design decision.

Other differences between tutors and games, such as how to provide assistance, are more difficult to resolve. Tutors typically provide step-level, teaching feedback, either directly or via hint messages (VanLehn 2006). Tutors also allow students to immediately correct their errors. This is in stark contrast to games. Games rarely provide knowledge-based feedback and, instead of allowing immediate error correction, games typically impose penalties for making mistakes such as decreasing health or death. These different design approaches to assistance lead to very different experiences. What would be considered undesirable floundering in a tutor is not uncommon in games.

Previous work on cognitive tutors suggests that immediate knowledge-based feedback should be the most effective approach (Corbet 2001). However other work on intelligent-novice feedback (Mathan 2005) and situational feedback (Nathan 1998) suggest that there may be cases in which the

situational feedback provided by games may be as good, or in some cases better than knowledge-based feedback. One of the most maddening gaps in previous research is that it gives us no hint as to whether the style of feedback provided by games is a necessary part of their allure or simply an educationally ineffective convention. For example, it's hard to image that a player would experience the satisfaction of beating a game if the game imposed no penalties, and immediately tells the player what to do when they are stuck. On the other hand, for a task with a high, fixed level of difficulty, the lack of assistance in a game might produce so much floundering that the game proves too difficult to be fun. The problem is at the core of the assistance dilemma (Koedinger & Aleven, 2007).

In addition, there is little empirical evidence to suggests that a game is the most effective way to teach policy reasoning. A review of the empirical research on games by Hays (2005, p. 6) concludes that "the empirical research on the effectiveness of instructional games is fragmented .. and plagued with methodological flaws. … [it] does not tell us whether to use a game for our specific instructional task … [and] … there is no evidence to indicate that games are the preferred instructional method…"

This leads to the fourth research question: *if we use a combination of games and tutors to teach deliberation, is it better to provide game-like or tutor-like assistance?*

With respect to the learning challenges and learning principles of the learning elements layer, we can see that previous research leaves us with more questions than answers. It tentatively predicts that bias will present learner challenges both during search and analysis. It also suggests that we might try using external representations and combining tutors with games, but makes no predictions as to whether these tactics will prove effective, and no guidance as to how to actually implement these approaches for policy reasoning. It seems that at the leaning elements layer, our work is cut out for us.

Layer 4 :: Instructional systems : Inquiry environments, tutors and games

We now move from low-level theories of learning elements to the design of instructional systems. Basic research on learning elements can constrain the design of systems and provide design hypotheses, but the generality and paucity of this research on learning elements leaves the design of a specific system underdetermined. Here we must look to examples of instructional systems designed for other domains in the hopes that they can provide clues for designing a deliberation tutor.

An instructional software system includes three sub-layers: (a) the delivery system by which students access the instructional system, (b) the inquiry environment in which problem solving takes place, and (c) assistance provided to the student during problem solving.

The delivery system: the Open Learning Initiative

Fortunately, previous work by the Open Learning Initiative (OLI), Thille (2008), solves the problem of delivery. OLI provides a web-based delivery system that allows instructors to deploy more or less arbitrary web-based software and static web pages. OLI offers general services for handling student registration and access, logging student actions, and recording student work. The delivery system is the only part of the learning environment platform for which previous work generalizes well to the

problem of teaching deliberation. From an academic perspective, it is ironic to note that this generalizable research comes out of practice rather than basic research.

The inquiry environment: Microworlds, recording tools, representation tools, and scaffolding

The second sub-layer of the instructional system layer is the inquiry environment in which problem solving takes place. Again, given the lack of work on inquiry environments for deliberation, we must look to other domains. A number of communities have developed relevant technologies including: Games For Change, inquiry learning environments, and computer supported argument visualization (CSAV) tools. Games For Change lists upwards of 60 games addressing a wide range of topics such as environment, health and poverty (Appendix C). Unfortunately, most of the games: (a) focus on teaching content knowledge rather than skills, (b) are designed based on entertainment rather than educational principles, and (c) are not evaluated, so it is difficult to claim that they will improve deliberation skills. Inquiry learning research has produced scores of learning environments and tools for teaching science and math (Appendix A), but not for deliberation. Computer supported argument visualization researchers have produced a number of representation tools, primarily using argument diagrams, that have been used to teach argument or to discuss deliberation topics (Appendix B), although few of these use causal diagrams. In other words, there is generally little research on how to improve the skills of deliberation that we are concerned with here.

None of these tools and environments provide us with anything remotely resembling a deliberation tutor from which we could add incremental improvements. But examining how the the tools and environments might relate to the deliberation framework allows us to abstract the basic contours of an inquiry environment for deliberation. Abstracting across these examples, we can see that inquiry environments provide four types of tools and mechanisms: (a) a microworld, (b) recording tools, (c) representation tools, and (d) process scaffolding. The **microworld** provides the raw information that is used to solve a problem, for example in science education, microworlds take one of three forms: a fixed data set such as data tables about the survival rate of Galapogos finches, a microworld in which observation can be made such as a ecological simulation of air pollution, or sensors that allow students to collect real world data. **Recording tools** allow students to record observations, annotate evidence, or archive their materials. **Representation tools** allow students to create charts, diagrams, maps and other external representations needed for problem solving. Representation tools can be partially automated, for instance a data analysis tool that takes a data set and produces a bar chart automates the construction of the representation. A modeling tool that takes a causal diagram and simulates how outcome variables change in response to interventions partially automates the process of interpretation. All the work on CSAV concerns representation tools. Finally, **process scaffolding**, helps students identify the goals and subgoals of inquiry.

Not coincidentally, these tools align with the deliberation framework: microworlds allow opportunities to practice search skills, recording tools support comprehension (and archive the products of problem solving), representation tools support construction, synthesis and interpretation, and goal states provide a direction for questioning. So while there may be little work on environments for deliberation, we at least know the *types* of tools that must be provided by a deliberation environment.

Assistance: Causal reasoning tutors and argumentation games

Once the inquiry environment has been developed, the next sub-layer of the tutoring platform defines the assistance provided. To inform the design of an assistance service that promotes both learning and motivation, I look first at the type of feedback provided by different causal reasoning tutors and second at the intelligent tutoring behavior of games.

Cognitive tutors for causal reasoning

Although there are no tutors for deliberation, there are a number of tutors that teach causal reasoning, including: Betty's Brain (Leelawong & Biswas, 2008), 20/20 (Masterman, 2005), VModel (Forbus, Carney, Sherin, & Ureel, 2005), SEEK (Graesser, Wiley, Goldman, O'Reilly, Jeon, & McDaniel, 2007), Sourcer's Apprentice (Britt & Aglinskas, 2002), and the Why System (Collins & Stevens, 1977). In Betty's Brain, students construct causal diagrams on global warming that represent the knowledge of a virtual student named Betty who 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. VModel allows students to construct causal diagrams and run simulations. Students using SEEK and Sourcer's Apprentice read evidence of varying reliability before writing essays on the causes of volcanic activity or the Panamanian Revolution. The Why System, a text-based Socratic tutor, asks students questions to help them to create a mental representation of the causal factors involved in growing rice.

Table 2.1

Scaffolding and Feedback on Deliberation Steps Provided by Causality Reasoning Tutors

Tutor	Question	Search	Comprehension	Evaluation	Synthesis	Construct	Interpret
Betty's Brain	-	Reliable	-	-	-	Betty's quizzes	Betty's explanations
VModel	-	-	-	-	-	Syntactic feedback	Predictions feedback
20/20	-	-	-	-	-	Model feedback	-
SEEK	-	Varying	Form	Credibility feedback	-	-	-
Sourcer's Apprent.	-	Varying	Form + Feedback	-	-	-	-
Why System	Socratic	-	-	-	-	Socratic	Socratic

With respect to deliberation framework, 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 and provides a hint button with search suggestions, whereas 20/20 and VModel do not provide information as part of the inquiry environment. Note that tutors could teach search using microworlds (Jonassen & Ionas, 2008) as in the Causality Lab (Scheines, Easterday, & Danks, 2007) which allows students to 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. Note however, that tutoring is possible; the Why System (which predates GUIs) asks Socratic questions to help the student create a mental representation of the causal system.

Tutors using causal diagrams provide feedback on **construction** in different ways. 20/20 provides the most explicit feedback by immediately comparing student's 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, e.g., 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.

While none of these tutors teaches policy reasoning per se, they do show that feedback can potentially be provided on most stages of deliberation. The limitation of previous research lies in the scope of problems tutored. Our goal is to teach students how to make decisions about policy problems. In these problems, students must synthesize information from conflicting sources to make decisions. Yet none of these tutoring systems teach all the steps of deliberation. Even the combined set of systems would not fully teach deliberation, because it would not provide assistance on synthesis. This limitation applies not only to policy, but also to science, history and argument, which all require synthesis of conflicting information.

Argumentation games

In addition to causal reasoning tutors, we can also look to a number of *argumentation games*. These are games in which players must either make arguments or support hypotheses based on evidence. The games include: *Argument Wars* (2010), *Advisor to the King* (Hastings, Britt, Sagarin, Durik, Kopp, 2009), *Crystal Island* (Mott & Lester, 2006), *Global Conflicts: Latin America* (Serious Games Interactive, 2008), *Phoenix Wright* (2005), *Resilient Planet* (The JASON Project, 2007), and *Scientopolis* (Nelson, Ketelhut, Schifter, 2009). In *Argument Wars*, students argue Supreme Court cases on issues like whether or not schools have the right to search student's lockers without a warrant. In *Advisor to the King*, students identify claims in different texts. In *Crystal Island* students explore a 3D island making observations about a mysterious outbreak which they must identify. In Mission 2 of *The Operation: Resilient Planet Game*, students explore an underwater environment collecting observations about turtle's habitat. In *Scientopolis* students gather data about sickness among different groups of sheep. In *Global Conflicts: Latin America* students interview residents of a Mexican village about the Maquiladoras and then confront the factory owner. In the entertainment game *Phoenix Wright*, students search for evidence via interviews and physical search, which they use

to cross-examine their opponent by submitting that evidence when it contradicts their opponent's claims during a court case.

Table 2.2

Feedback Provided by Causal Reasoning Tutors and Argument Games on Steps of Deliberation

System	Search	Comprehension	Evaluation	Construction	Interpretation	Argument
Tutors						
Betty's Brain	+			+	+	+
20/20				+		
VModel				+	+	
SEEK	+	+	+			
Sourcer's Appr.		+				
Why system				+	+	?
Games						
Argument Wars						+
Advisor to the King		+				
Crystal Island						+
Global Conflicts						+
Phoenix Wright						+
Resilient Planet	+	+				+
Scientopolis						+

Most of the games (and tutors) provide little support for search. In *Argument Wars* students "search" freely for different argument supports by drawing and discarding cards at will; likewise in *Crystal Island* and *Scientopolis*, students roam a 3D world freely with no guidance. *Global Conflicts: Latin America*, and *Phoenix Wright* involve a more complicated search space in which players must ask specific characters certain questions before other types of evidence can be found or before certain answers are provided. In *Phoenix Wright* the game requires the player to find all the required evidence before they can proceed to the courtroom; whereas in *Global Conflicts: Latin America* the student may confront their opponent even if they do not have enough evidence to win. *Betty's Brain* and *The Operation: Resilient Planet Game* provide a certain amount of scaffolding: in *Betty's Brain* key terms that should be diagrammed are highlighted, and in *Resilient Planet* objects of interest are sometimes highlighted with a visual marker. Only *SEEK* provides explicit help searching for web pages via a hint button (although it is not clear whether or not these suggestions are static.)

Comprehension is also rarely supported. *Advisor to the King* is exclusively about comprehending claims, so it gives correctness feedback after the student attempts to identify a claim. *Resilient Planet* has some support for comprehension via a taxonomy tool – when the student photographs an animal, they are asked a series of yes/no questions, e.g., "Are the scales blue?" until they have identified the animal photographed. In this way the student's attention is drawn to specific features of the animal.

The games also ignore evaluation, comprehension, construction, and interpretation, because they do not ask students to assess the value of the information they have found or create explicit representations of the evidence.

However, all the games provide feedback on students' arguments, since this is how students either win or lose the game. In *Argument Wars*, whenever a student proposes that one of their cards supports an argument or objects to an opponent's card, the student is given immediate correctness feedback. *Phoenix Wright* and *Global Conflicts* also provide correctness feedback when the player presents a piece of evidence to contradict the opponent's claim. *Resilient Planet* has a similar argument mechanic where the team leader asks the student to choose between two statements, e.g., whether a certain kind of turtle prefers deep or shallow water, and to provide an evidence card that supports the student's claim. In *Crystal Island*, students use their observations to fill out a fact sheet, which, if correct, identifies the virus responsible for the outbreak. It should be noted the "arguments" created in these games do not even begin to approach the level of sophistication that we expect from a high school debate – in reality these games typically ask a multiple choice question about whether a piece of evidence supports a claim. This is only a subset of argumentation, albeit an important one.

In comparing these games to the causal reasoning tutors, we see two general differences: the argument games tend to involve tasks of a larger scope than the tutors, in that students must both search for evidence and use that evidence to make an argument. Along these lines, much of the development effort in argument games seems to focus on creating a compelling (often 3D) world for the student to explore and from which to gather observations. Unfortunately, this focus on the microworld seems to be at the expense of teaching analysis, and it is not clear what direct educational benefit is gained from running around a virtual environment. The causal reasoning tutors on the other hand seem to focus on smaller subtasks, but provide a greater amount of feedback than the argument games.

These examples show that we can design games based on argumentation. This includes simulated debates as in Phoenix Wright and limited cross-examination as in Global Conflicts. They suggest tantalizing possibilities for how an intelligent deliberation tutor might increase motivation. However, this previous work has many limitations. First, the argumentation tasks in these games pale in comparison to the task of creating or debating a real argument. Second, if we follow the conventions of argumentation games for providing assistance, then we will provide far less assistance than provided by a typical causal reasoning tutor. This leads to the dilemma or whether adding more assistance will increase learning, decrease interest, or possibly both. Previous research gives us little insight in how to design an educational game on deliberation to maximize both learning and motivation. The differences between argumentation game and deliberation tutors also creates a third dilemma. Much of the development effort in argument games seems to focus on creating a compelling (often 3D) world for the student to explore and from which to gather observations. Unfortunately, this focus on the microworld seems to be at the expense of teaching analysis, and it is not clear what direct educational benefit is gained from running around a virtual environment. It is not clear whether these 3D worlds are necessary for increasing motivation, whether search should be a primary focus of instruction, or whether navigating a 3D environment is an effective use of instructional time.

This leads us to the fifth research question: *how can we develop an instructional system that combines a game-like inquiry environment with an intelligent tutoring system to teach deliberation?*

Layer 5 :: Curriculum : Civic education

Although the design of a full curriculum for civic engagement is beyond the scope of this dissertation, the intelligent tutoring system at the core of this work is designed with specific civic curricula in mind. There are now a number of curricula that rely upon or are designed to teach civic skills during the solving of real community development problems. All of these curricula can benefit from automated instruction in deliberation. These curricula include: Carnegie Mellon University's Technology Consulting in the Community, the Peace Corps, AP Government, and Deliberative Polling. 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 projects impact the mission of their non-profit clients. Similarly, the 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. 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). In keeping with the tutoring design principle focused on classroom use from the start (Koedinger & Corbett, 2006), the instructional system described here is intended for use in these curricula with relatively minor modification.

What does research tell us?

What does research tell us about designing a deliberation tutor? Instruction in deliberation has received little attention. This chapter has performed due diligence by attempting to generalize the findings of work in several disciplines and domains to deliberation. Specifically it proposed a learning environment platform and a deliberation framework that decompose the problem into smaller sub-problems. It then analyzed previous research to identify which sub-problems have potentially been solved and which must still be answered. Unfortunately, examination of a relatively broad body of the most relevant previous research leaves us with more questions than answers.

There is a paucity of studies on policy reasoning tasks, none of which examine a whole policy problem. The deliberation framework (the first contribution of this dissertation) had to be constructed through an analytical process in order to account for how a reasoner might solve a policy problem. It was then supported by identifying empirical studies on related skills used in other domains. The cognitive psychology research on learning elements provides tentative warnings about bias, but as will be seen in later chapters does not accurately specify where bias occurs. The related hypothetical learning tactics also provide little guidance in terms of instructional design, nor does this research guarantee that these learning tactics will prove effective. Looking to related instructional systems, we are again provided with little guidance in how to construct an inquiry environment for deliberation or how to provide assistance. However we can take heart that assistance can be provided on many of the skills of causal reasoning, and that entertainment games have been created using argumentation mechanics.

Previous work leaves us with a number of components at which to "hack away". Specifically, the dissertation must address five research questions:

1. *What cognitive model best describes deliberative skill?* to which a preliminary answer has been provided.
2. *What learning challenges do students face in acquiring this cognitive model?,* which will be addressed in Chapter 3.
3. *Can causal diagrams help students overcome difficulties in policy reasoning?,* which will be addressed in Chapter 4.
4. If we use a combination of games and tutors to teach deliberation, *is it better to provide game-like or tutor-like assistance?* which is a scaling-down question that can only be addressed after an instructional system has been designed. This question will be addressed in Chapter 6.
5. *How can we develop an instructional system that combines a game-like inquiry environment with an intelligent tutoring system to teach deliberation?* which will be addressed in Chapter 5.

3. Localizing reasoning difficulties in synthesis

Summary. Deliberation requires students to choose policy positions based on evidence, yet confirmation bias prevents them from doing so. Here we investigate the learning challenge of bias by asking: *where does bias occur during the search for and analysis of evidence in a policy reasoning task?* In this on-line, laboratory study, 60 university students played a game in which they chose which of four policies would increase school performance. Students were randomly assigned to either search for evidence using a Google-like environment, or read all available evidence sequentially. The evidence presented to each group was manipulated to be congruent or incongruent with two of the student's prior beliefs. The study used a two-group design, with search (Google / sequential) as between-subjects manipulation, and congruence of evidence with the student's belief as a within-subjects manipulation. The study measured students' evidence-based recommendations, their change in beliefs, and their recall of the evidence. The study found that students did not cherry-pick evidence nor discount disconfirming evidence. However, students' extreme confidence in their initial beliefs usually prevented them from changing position, and students mistakenly recalled the evidence as confirming their beliefs. These results suggest that deliberation tutors should focus on evidence synthesis and making recommendations based on explicit evidence.

With the cognitive model of deliberation from Chapter 2 in hand, we begin our empirical investigation at the learning elements layer, in order to determine the learning challenges that students face when trying to acquire deliberative skill (Figure 3.1). This chapter will address the second research question: *what learning challenges do students face in acquiring the cognitive model of deliberation*? Specifically, in which of the steps of policy reasoning does bias occur?

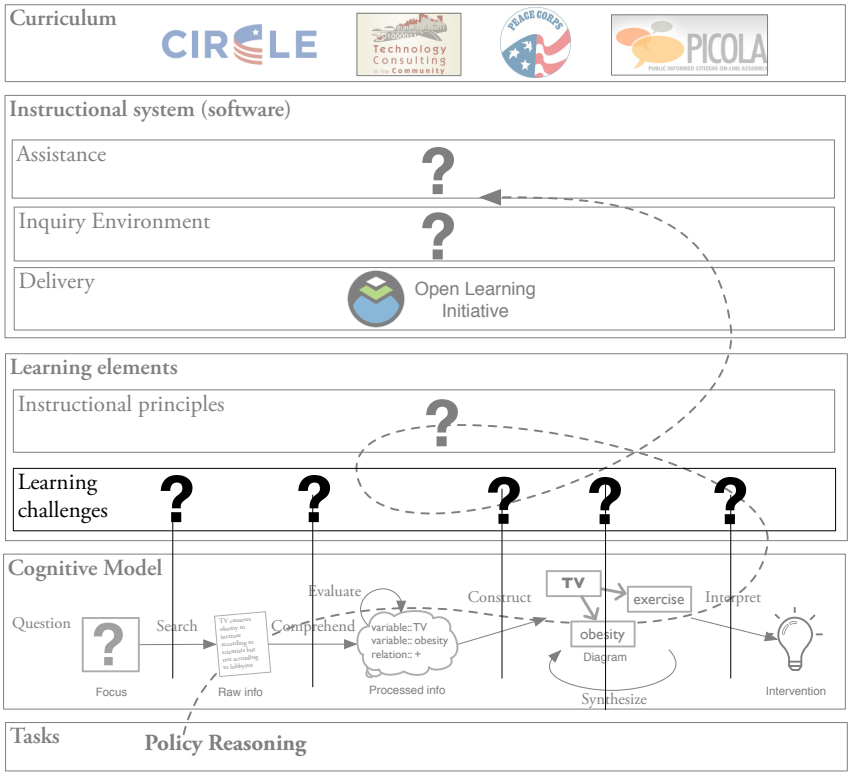


Figure 3.1. Chapter 3 identifies learning challenges by localizing where confirmation bias occurs in the steps of deliberation.

The study presented here tests where bias occurs during a policy reasoning task requiring search and analysis of evidence about school performance. At the end of the task, reasoners must decide which of four policies for increasing school performance should be recommended. Although the task will be relatively simple, its scope is much larger than in many of the previous studies described in Chapter 2, some of which do not present students with any evidence at all. This slowly moves us toward a more authentic task for a policy tutor (Edelson & Reiser, 2006). Despite the limited scope of previous work, taken together, these studies predict that bias will arise both during the search and analysis of evidence.

This study examined the effects of: (a) the ***congruency of evidence*** with students' prior beliefs, and (b) students' ***search*** for evidence on their ability to make ***evidence-based recommendations***. For example, if a student believes that decreasing classroom size doesn't increase school performance, then he might not recommend decreasing classroom size even if he is given evidence to the contrary. To examine *congruency of evidence* with belief, the study provided evidence mostly congruent with one of the student's beliefs about a policy (e.g., decreasing class size), and provided evidence mostly incongruent with one of the student's beliefs about a second policy (e.g., requiring teachers to have masters degrees). If confirmation bias is an issue, then we should see the student agree with evidence when it is congruent with her prior beliefs about the first policy, and disagree with the evidence when it is incongruent with her prior beliefs about the second policy. To examine *search,* the study compared a free search group in which students searched for evidence in a simulated Google environment (the *Google* group) to a sequential presentation group in which students read every piece of information in a fixed order specifying their beliefs after each piece of evidence (the *sequential* group). If confirmation bias is a problem during search, then we should see differences between how the groups respond to evidence. Specifically, we expect Google students to search for evidence that will be congruent with their beliefs. The study measured the ***evidence*** read by each student, students' ***confidence shifts*** in their beliefs about the policies after reading the evidence, and students' ***recall of the evidence*** analyzed**.**

Based on the consistent pattern of confirmation bias in previous research, I predicted bias at all stages of processing:

H1: Biased search. Google students will search for reports in a biased manner, i.e., search for a higher proportion of reports congruent with their beliefs.
H2: Biased evaluation. Students will shift their beliefs more in response to evidence that is congruent with their belief than to evidence that is incongruent.
H3: Biased synthesis. Students will overestimate the amount of congruent evidence read.
H4: Biased decision. Students will make recommendations more consistent with their beliefs than with the evidence.

Method

Population and setting

Sixty university students were recruited through an on-line subject database to participate in the study. Participants had a median age of 21 years, 95% were bilingual or native English speakers, and 92% used the internet at least once a day. Participants completed the study over the internet.

Procedure

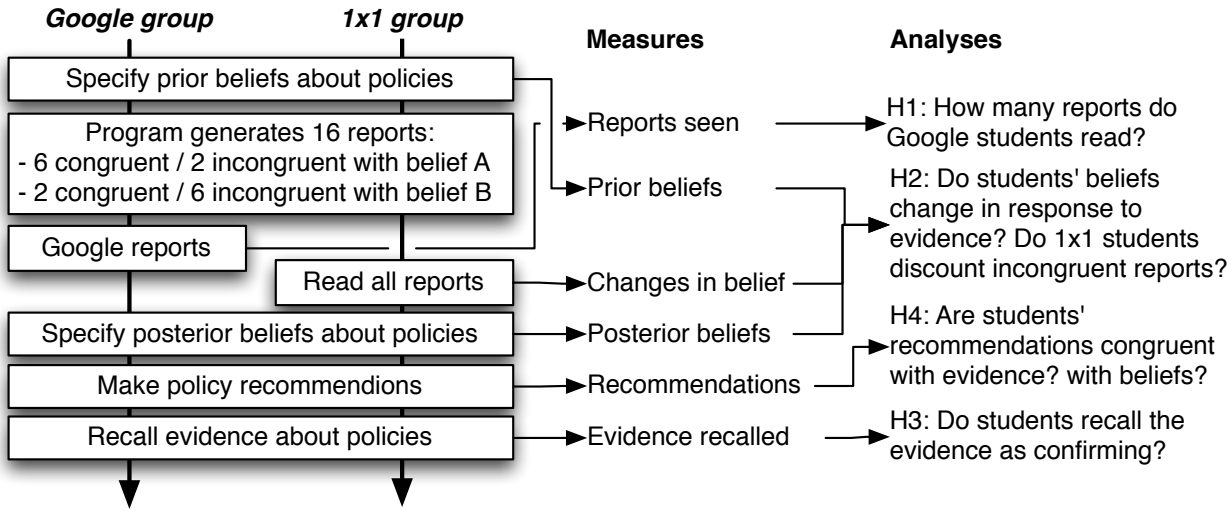


Figure 3.2. Experimental procedure, measures and analyses. The experiment used a 2-group design, randomly assigning students to either the sequential or Google groups. A within-subjects variable manipulated whether the two sets of evidence were congruent or incongruent with the student's prior beliefs.

Students played a computer game in which they assumed the role of policy analysts. Their goal was to determine whether four different policies: reducing class size, increasing teacher qualifications, increasing funding, or providing vouchers, would increase school performance. At the beginning of the game, students specified their prior beliefs about each policy: whether the policy would have a positive, negative, or negligible effect on school performance, and their certainty in their belief (Figure 3.3.)

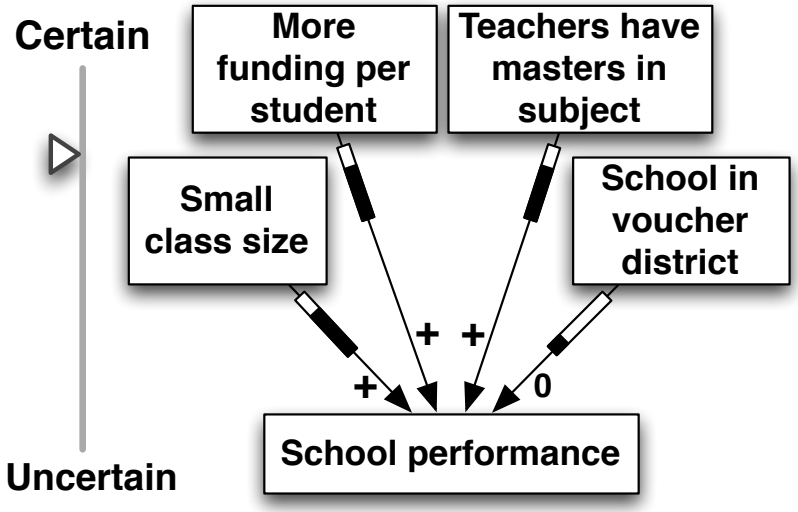


Figure 3.3. Graph representing the user interface students used to specify their beliefs about the effects of interventions on school performance. A student specified his belief about each policy by toggling a +/0/- button on a causal arrow and moving a slider to indicate his certainty in that belief.

In both groups, the game identified the two policies about which the student had the strongest beliefs and generated two sets of evidence in the form of one-paragraph descriptions mimicking newspaper reports (Figure 3.4). One set of 8 reports was *mostly* congruent (6 congruent, 2 incongruent) with one of the student's beliefs (e.g., about class size), and a second set of 8 reports was *mostly* incongruent with a second policy belief (e.g., about teacher qualifications). Half the reports summarized observational studies, and half case studies (in order to test whether students were sensitive to evidence type). The game then randomly assigned students to either the *Google* group which searched freely for reports, or *sequential* group which read every report and specified the change in their beliefs after each report.

Example report:
Dr. Jones, a professor of educational policy at Harvard University, studied high schools from 12 different states on many dimensions including class size. The schools were evaluated on how well their students performed on the NAEP test of mathematics. Mr. Jones found that schools with smaller classes performed no differently than schools with larger classes on the NAEP. When asked about the implications of this research, Dr. Jones implied that more work like this needs to be done in order to fix America's schools.

Figure 3.4. Example report on the effect of an intervention on school performance. The report summarizes an observational study that shows no effect of smaller class size on school performance which would be incongruent with a prior belief that smaller class size increases performance.

Google students used a fake Google interface to search simulated web pages generated by the game to find the reports (Figure 3.5). Students could not change the search query, but they could choose which search results to examine and which reports on which sites to read. After clicking on one of the search results, Google students saw a home page stating the organization's biases (Figure 3.5, bottom-left). If the student were to exhibit the sort of bias found in previous studies, he might decide not to search the site further if the organization's bias is incongruent with the his prior beliefs. If the student decided to search the site further, he would find a list of reports (Figure 3.5, bottom-right) leading to an individual report like that in Figure 3.4.

The sequential students did not have a Google interface and were simply presented with each of the 16 reports one at a time.

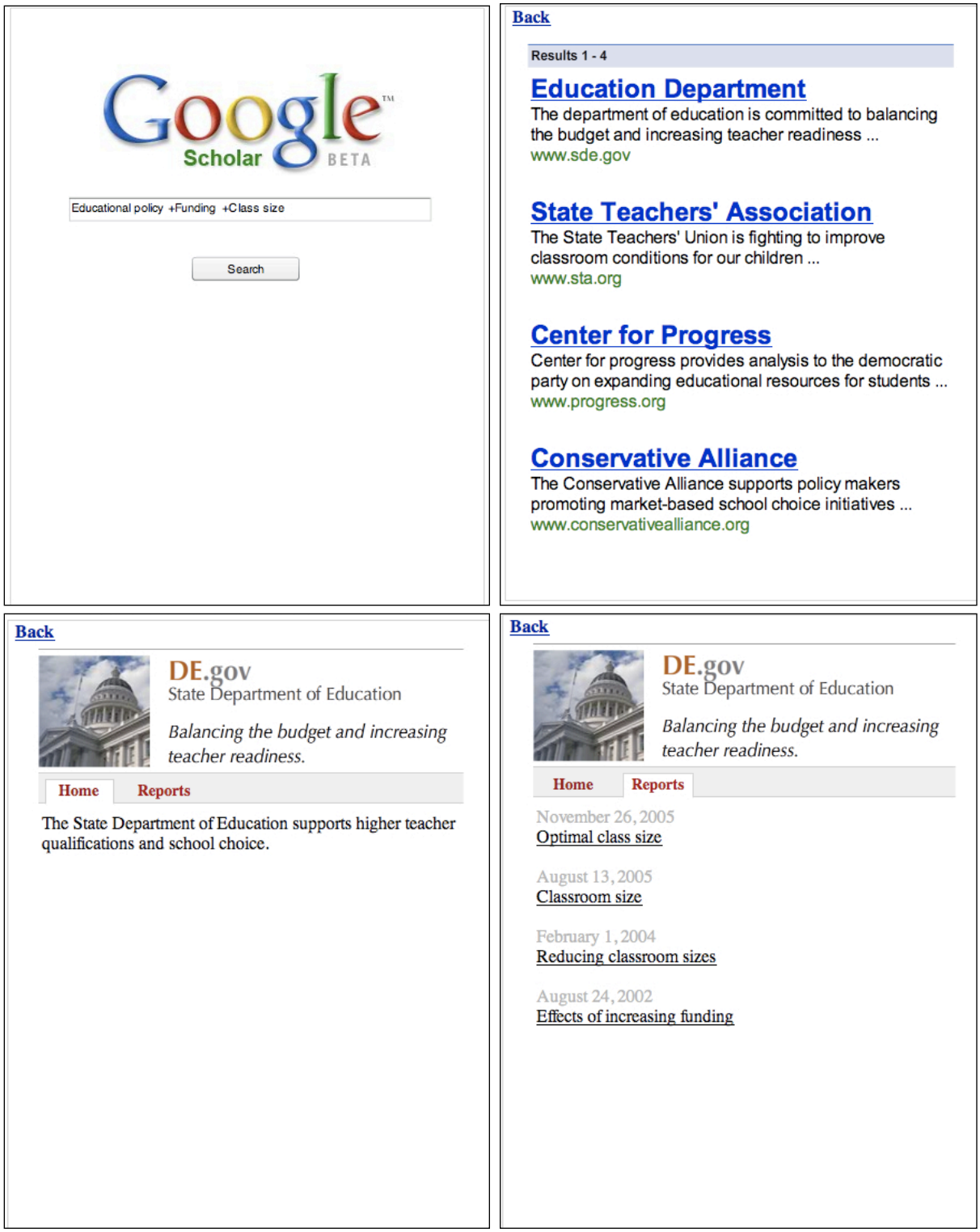


Figure 3.5. Screen shots of the Google group's search interface.

After reading the reports, all students specified their final beliefs about how each policy would affect school performance, made their recommendations about which policies should be implemented (yes/

no), and specified how much evidence they thought they read about each policy (number of reports seen about each policy and the proportion of reports indicating the policy would work).

Results

H1: Biased Search

To test the first prediction that Google students would search in a biased manner, the first analysis compared the reports read by each group and the proportion of the reports read that were congruent with the students' prior beliefs.

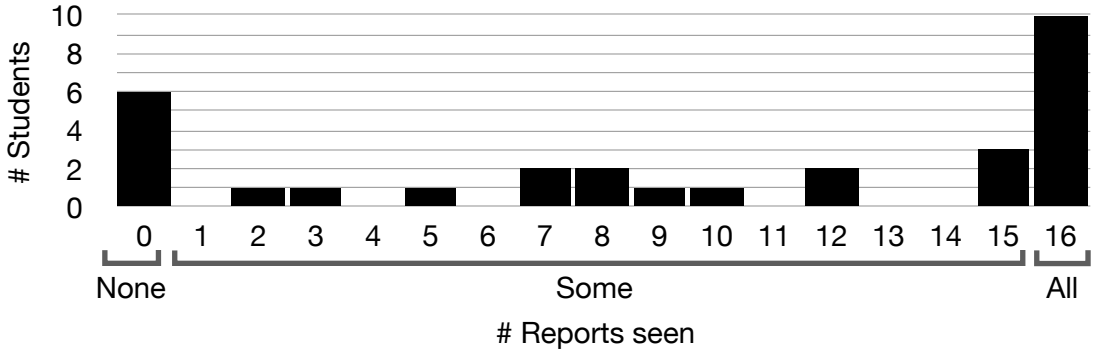


Figure 3.6. Number of students in Google group who searched for the given number of reports.

Figure 3.6 shows the raw number of reports seen by each Google student. It shows that about two thirds of the Google students did not read all the reports. Six Google students read none of the reports, 14 read some (1-15) reports, and 10 read all reports.

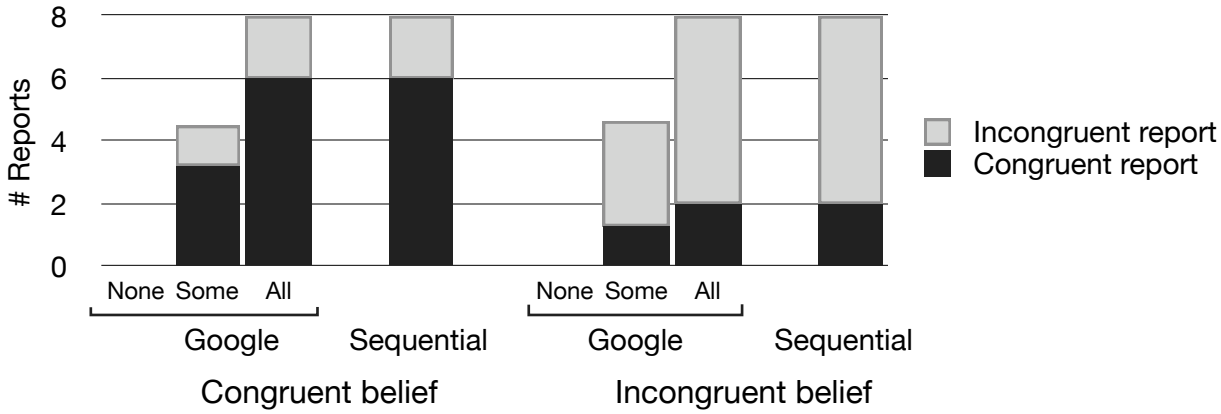


Figure 3.7. Number of congruent and incongruent reports read by each group. The graph shows the number of congruent (dark gray) and incongruent (light gray) reports read by the sequential students and the Google students who read all, some, or no reports.

Figure 3.7 shows the proportion of congruent/incongruent evidence seen for two of the students' beliefs, the belief for which the experiment provided mostly congruent evidence (Figure 3.7, left), and the belief for which the experiment provided mostly incongruent evidence (Figure 3.7, right). Figure 3.7 displays separately the Google students who search for none, some, or all reports.

For the belief mostly congruent with the evidence, the 30 students in the sequential group read virtually the same proportion of confirming evidence (M = 0.75, SD = 0) as the 24 Google students who read some or all reports (M = 0.71, SD = 0.22), t(23) = 0.93, p > .36. For the belief mostly incongruent with evidence, the 30 students in the sequential group read virtually the same proportion of confirming evidence (M = 0.25, SD = 0) as the 23 Google students who read some or all reports (M = 0.26, SD = 0.26), t(22) = 0.50, p > .62.

In other words, the Google students did not always search for all evidence, but they did not appear to search in a biased manner. This finding fails to confirm the first prediction that Google students would selectively search for evidence congruent with their prior beliefs.

H2: Biased evaluation

The second hypothesis predicted that students will shift their confidence in their beliefs more when given evidence congruent, rather than incongruent, with their beliefs. In the extreme case, this predicts that students will increase their confidence in their original beliefs and completely ignore contradictory evidence. To test this hypothesis, the game logged students' beliefs about each of the four policies before and after reading the evidence. Analysis of the total shift in belief by each group (Figure 3.8) did not show the sort of belief polarization found in other policy reasoning studies. In fact, even for the policy belief for which mostly congruent evidence was provided, students decreased their confidence in their original belief. Recall that, for the belief for which they were provided congruent evidence, 2 of the 8 reports conflicted with the student's belief. This means that f students' prior beliefs are so extreme that the expect to see only 0 or 1 incongruent report, then normatively, they should decrease their confidence in their original beliefs.

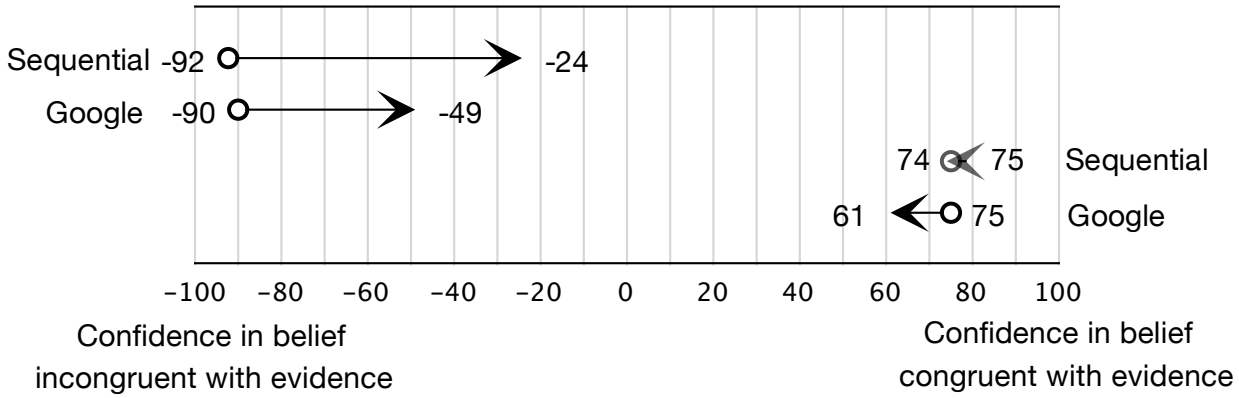


Figure 3.8. Shifts in confidence by each group after reading mostly congruent or incongruent evidence.

The 30 sequential students decreased their confidence in the incongruent belief by 67% (SD = 71%) and in the congruent belief by 1% (SD = 40%). The 30 Google students decreased their confidence in the incongruent belief by 41% (SD = 55%) and in the congruent belief by 19% (SD = 59%). A linear mixed model regressed students' shifts in confidence on whether the student was in the Google or sequential group, and whether or not the student's initial belief was congruent with the evidence, with student as a random effect. Students decreased their confidence (b =) 65% more when the evidence was incongruent with their belief, t(59) = 6.18, p < .0001. Google students did not shift their confidence in the direction of the evidence as much as sequential students, (b = 22%), t(58) =

-2.13, p < .04. According to this model, students do in fact change their beliefs more when their beliefs are incongruent with the evidence, as they should.

Lacking a normative psychological theory of belief updating, we can (imperfectly) define confidence shift as *positive* when it is toward the belief supported by evidence, and *negative* when it is away from the belief supported by evidence. Given this definition, Figure 3.8 shows that students' confidence shifted more in response to incongruent evidence than to congruent evidence, and that sequential students shifted their confidence more normatively (in the direction of the majority of evidence) than Google students. Note however that an alternate model of *absolute* confidence shift showed no significant difference between Google and sequential students.

In the sequential condition, the game also solicited students' beliefs after reading each report. Analysis of the sequential students' shifts in confidence after each report showed that on average, students shifted in the correct direction by 12% after reading each report (*SD* = 38%), shifting in the wrong direction only 7% of the time. Students also responded more to incongruent reports (*M* = 17%, *SD* = 44%) than to congruent reports (*M* = 8%, *SD* = 33%) which a linear regression analysis indicated was a significant difference (b = 8.6%, p < .15) accounting for 1% of the variance.

Taken together these results fail to confirm the hypothesis that students will discount disconfirming evidence. However, students did not behave as normatively as one might desire. Their initially high confidence in their prior beliefs meant that their shift in confidence did not always result in a qualitative change in belief. For example, they did not tend to shift from a belief that smaller classes increase performance to a belief that class size doesn't matter. Furthermore, sequential students' data show that they did not shift their confidence more in response to observational studies than to case studies. This indicates that student were not sensitive to the type of evidence presented.

H3: Biased synthesis

To test the third hypothesis that students recall the evidence as more congruent with their beliefs than what they actually read, I asked students to specify how many reports they read about each policy, and the percentage of reports that indicated that the policy would work.

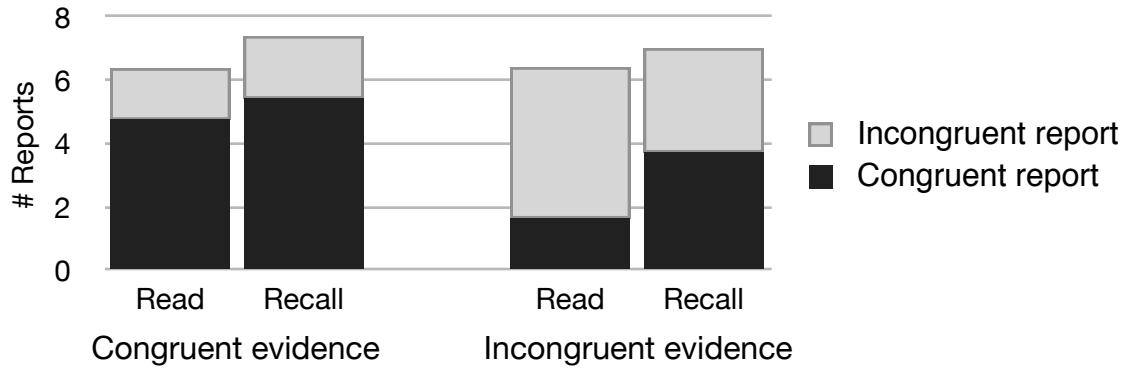


Figure 3.9. Number of congruent/incongruent reports read and recalled for the belief congruent with most of the evidence and the belief incongruent with most of the evidence.

Figure 3.9 shows the ratio of congruent and incongruent reports read and recalled. Here the Google group and sequential group are combined due to the lack of significant differences between the two

groups. Recall that each group could potentially see 6 congruent and 2 incongruent reports about one policy, and 2 congruent and 6 incongruent reports about a second policy, but that Google students did not search for every report which lowered the average number of reports seen by both groups. Across groups, we see that students saw on average 4.7 congruent reports and 1.6 incongruent reports about the prior policy belief congruent with the majority of the evidence (Figure 3.9, first column). We also see that students saw 1.6 congruent reports and 4.8 incongruent reports about the prior policy belief incongruent with the majority of the evidence (Figure 3.9, third column). The third hypothesis asks how well they recalled what they read.

Students who read at least one piece of evidence about both the congruent and incongruent policy (n = 53) recalled the evidence about the congruent policy accurately (M = -2%, SD = 28%), but recalled the evidence about the incongruent policy as far more confirming than what they actually read (M = 26%, SD = 26%). A t-test showed that the difference (M = 28%, SD = 35%) was significant, $t(52) = 5.9$, $p < .0000003$.

These results show that when the majority of evidence read is congruent with students' prior beliefs, they recall the number of congruent and incongruent reports seen quite accurately. However, when the majority of evidence read is incongruent with students' prior beliefs, they not only recall the evidence inaccurately, they recall that the evidence was mostly congruent with their prior belief (even though it was not). These results support the hypothesis that students synthesize evidence in a biased manner.

H4: Biased decision

To test the fourth hypothesis that students make recommendations more congruent with their beliefs than with evidence available (or read), I measured students' prior and posterior beliefs and whether they recommended the given policy.

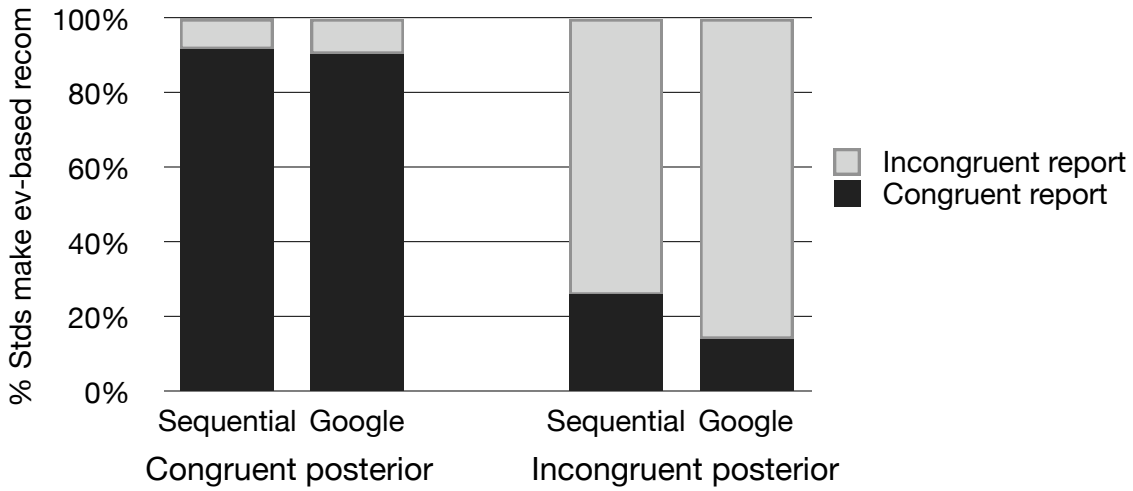


Figure 3.10. Evidence-based recommendations for congruent and incongruent posterior beliefs.

Students make recommendations congruent with the evidence 91% of the time when their posterior belief is congruent with the evidence (n = 69), but only 20% of the time when their posterior belief is incongruent with the evidence (n = 51), $t(85) = 10.9$, $p < 2.2e\text{-}16$. This shows that students'

recommendations are more consistent with their posterior beliefs than they are with the evidence, supporting the hypothesis.

Path Analysis

To understand the relation between beliefs, evidence, and recommendations, I analyzed the relations between: (a) the *search condition*, sequential or Google, (b) the student's *prior belief* about whether the policy had a causal effect, e.g., that they were 50% certain that the policy had a positive effect on school performance before reading the evidence, (c) the percentage of *evidence available* (reports) indicating the policy had a causal effect, (d) the percentage of *evidence read* by the student indicating the policy had a causal effect, (e) the *evidence recalled*, (f) the student's *posterior belief*, and (g) whether the student *recommended* the policy.

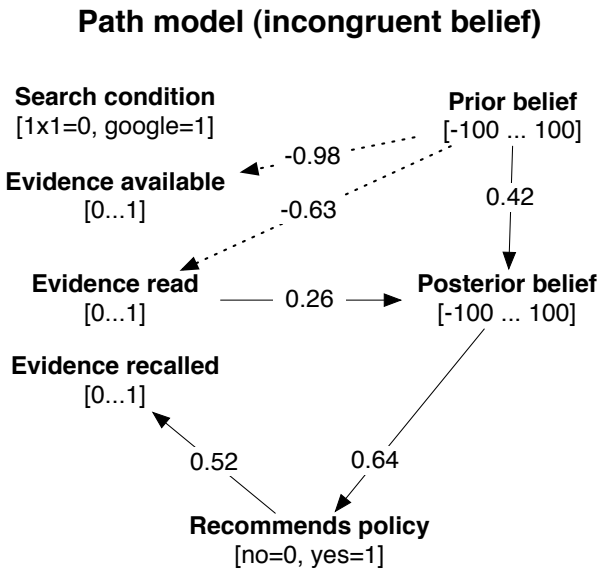


Figure 3.11. Path model of the effects of evidence and belief on recommendations and evidence recalled.

Table 3.1

Path Model Correlations for Policy Beliefs Incongruent with the Evidence (n = 60)

	Search	*Prior*	*Evidence*	*Ev. read*	*Ev. recall*	*Post*	*Recommend*	Mean	SD
Search	1.000							0.50	0.50
Prior	-0.091	1.000						81.83	41.54
Evidence	0.076	-0.976 ***	1.000						
Ev. read	0.250 *	-0.628 ***	0.616 ***	1.000				0.30	0.14
Ev. recall	0.006	0.165	-0.163	0.145	1.000			0.53	0.28
Post	0.132	0.418 ***	-0.411 ***	-0.101	0.384 **	1.000		26.83	69.46
Recom.	0.141	0.352 **	-0.324 **	-0.007	0.523 ***	0.645 ***	1.000	0.67	0.48

*$p < .05$ **$p < .01$. ***$p < .001$.

I used the Tetrad program (Tetrad 2008; Spirtes, Glymour, & Scheines, 2000) to search among the 221 possible path analytic models consistent with the correlations between the variables (Table 3.1)

and the knowledge that search condition and prior beliefs were set before evidence became available, which preceded the reading of the evidence, which preceded measurement of the evidence recalled, posterior beliefs, and recommendations. The path model (Figure 3.11) suggests that students' prior beliefs influence their posterior beliefs, which influence their recommendations. While evidence also influences students' posterior beliefs, it has a weaker effect than prior beliefs, and students' recall of the evidence merely rationalizes (rather than causes) their final recommendations. A chi-squared test of the deviance of the path model from the observed values (where larger p-values indicate better fit) showed that the predictions of the model did not differ significantly from the observed values, χ^2 (15, n = 60) = 21.78, p > .11.

Discussion

At first glance, the results seem contradictory and only partially consistent with previous literature: no bias was found in students' search or response to evidence, yet the evidence seemed to have little impact on students' final recommendations. The overall result that the evidence has little effect on students' recommendations is consistent with the results of previous work. However, by observing steps in problem-solving not recorded in previous studies, the study also showed that on the steps of search, and evaluation, students respond in a less biased manner than previous work might suggest.

Students appear to respond rationally to any given piece of evidence, increasing their confidence in response to congruent information and decreasing their confidence in response to incongruent information in a relatively symmetric way. However, because the strength of their prior beliefs is so high, each individual piece of evidence affects the prior belief only slightly. In other words, the evidence is like drips of water on the stone of belief. Furthermore, if students' beliefs represent only a single overall *impression* (Lodge, McGraw, & Stroh, 1989; Kim, Taber, & Lodge, 2008; Redlawsky, 2002) and do not catalog the evidence read, then when asked to recall information, students can only recreate an answer based on their overall confidence. Students do not so much process information in a biased manner as begin from an extreme position. But because they have an inaccurate picture of how the mass of evidence read compares to belief (biased synthesis), they do not recognize the inconsistency of their beliefs with the evidence.

Because we lack a normative theory for belief updating and students are unable to articulate the evidence supporting their initial beliefs, we cannot say whether students' initially high certainty is warranted. But we *can* say that their synthesis of the evidence is biased and inaccurate.

Implications for policy tutoring

These findings have several implications for developing a policy tutor:

1. A policy tutor should focus first on evidence synthesis (as opposed to search or comprehension). While other research has shown as many problems with search as with analysis, on this task students seem to have the greatest difficulty with synthesizing evidence.

2. Provide external representations of evidence that highlight where the mass of evidence contradicts belief. The study shows that students do not possess a clear picture of how the bulk of evidence supports or contradicts a particular causal claim. The first step in tutoring must be to provide

students with the skills and tools to recognize when evidence supports/contradicts a claim. External representations (e.g., diagrams, equations, etc.) are one approach to accomplish this.

3. Tutor an *explicit evidence* epistemic rule. Even if students recognize that the majority of evidence provided contradicts their belief, the evidence may still not be enough to change their belief. To improve performance on this task, the tutor must emphasize that students are not to make recommendations based on belief, but on evidence that they can explicitly cite. We might hope that over time, this will lead to beliefs that are more susceptible to evidence.

4. Using causal diagrams to improve policy reasoning

Summary. Novice policy reasoners face many learning challenges when solving policy problems like: *what should we do about global warming?* These problems 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. Diagrams might allow us to address these issues. As a first step toward building a policy deliberation tutor, I 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.

Chapter 3 examined the ways in which bias created *learning challenges* in the search and analysis phases of solving policy problems. We now move upward in the *learning elements* layer from *learning challenges* to *instructional principles* to ask the third research question: *can causal diagrams help students overcome difficulties in policy reasoning?* (Figure 4.1).

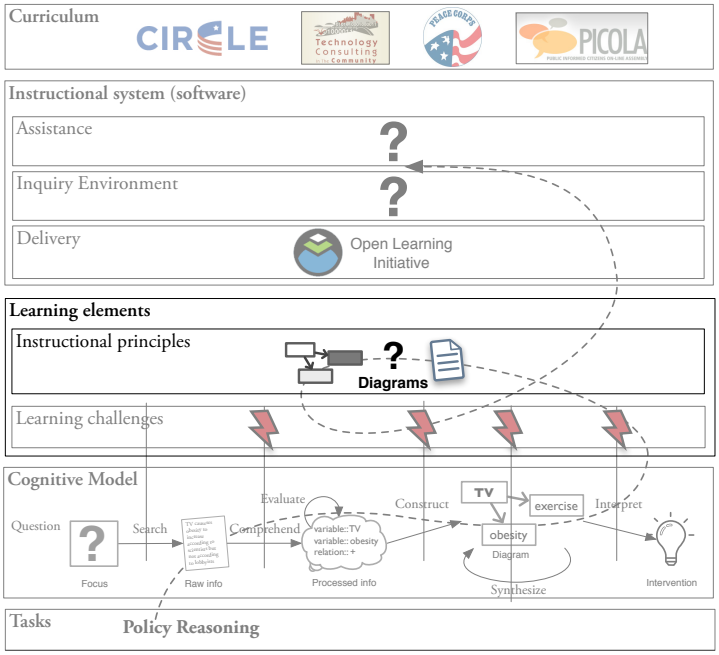


Figure 4.1. Chapter 4 tests the instructional principle: whether causal diagrams can be used to overcome the learning challenges in synthesis identified in Chapter 3.

As discussed in Chapter 2, there are three possible reasons for using causal diagrams in a deliberation tutor: (a) they *might* improve performance, (b) they *might* increase learning, and (c) they *do* provide a formal representation of the information that makes the problem less ill-defined and allows intelligent tutors to understand the students' beliefs. Causal diagrams might help to overcome some policy reasoning difficulties in the same way that equations help us solve algebra problems. External representations can relieve the memory burden required when using multiple pieces of information to solve a problem. In the case of diagrams, they may also also allow us to use our visual processing abilities to make cognitive inferences, i.e., to *see* an answer to the problem by looking at the diagram. The benefits of diagrams do not come "for free" however. Learning to use a certain type of diagram can impose its own set of learning challenges. Diagram users must both learn how to construct the representation and how to read its symbols. As we investigate the effectiveness of diagrams as an instructional tactic, we will also investigate the learning challenges by trying to identify barriers students face in using diagrams. Causal diagrams may also help students learn to solve policy problems if they help students gain knowledge and skills that can be used later in absence of the diagram, for example, the ability to encode causal information. The fact that causal diagrams better define policy problems and provide a machine-readable representation of the students' beliefs might be sufficient justification for using causal diagrams in a deliberation tutor. However, if diagrams increase either learning or performance, then teaching causal diagramming becomes a learning goal in its own right rather than just a technical necessity.

The empirical investigations presented in this chapter address the first two possible reasons for using causal diagrams, i.e., do they improve performance and learning? The studies also attempt to better understand the learning challenges associated with acquiring causal diagramming skill.

Research question: Constructing and interpreting causal diagrams to learn deliberation

To determine: (a) whether causal diagrams improve deliberation compared to text, (b) whether diagram construction promotes learning, and (c) the learning difficulties associated with constructing and interpreting causal diagrams, I 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 2.2, in Chapter 2), our study used three levels of external representation: text only, in which students solve problems unguided by a causal diagram, text plus a diagramming 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 or learning, because the learning challenges of constructing and interpreting 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 and learning, because constructing a diagram teaches one to better encode causal 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, I tested students on a posttest where diagrams or diagramming tools were provided and a transfer test where only text was provided. This allowed us to separate the effects on learning and performance.

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

STUDY 1: TEXT, DIAGRAMS, AND TOOLS

Method

Participants

Sixty-four 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.

Task

In this study, students were asked to read short policy briefs containing multiple causal claims from different sources like that in Figure 4.2.

> 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.

Figure 4.2. A text describing multiple causal claims from different sources about a policy topic.

In some cases, students were only given the text. In other cases, students were given a correct diagrammatic representation of the text like that in Figure 4.3 that appeared immediately below the text.

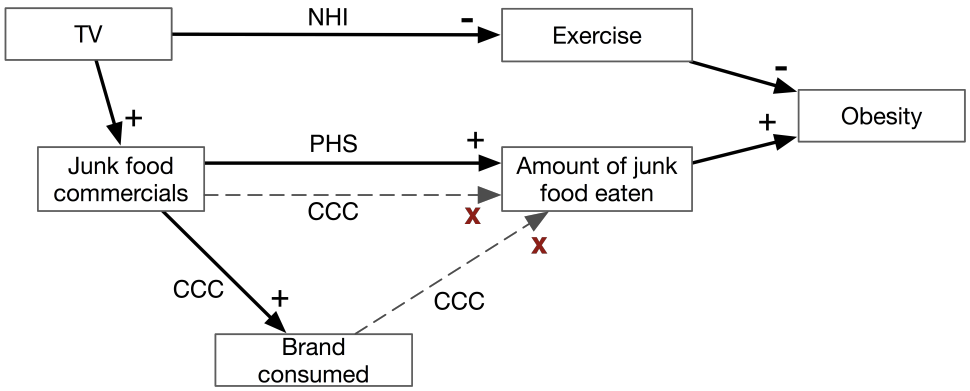


Figure 4.3. A correct diagrammatic representation showing all of the claims made my different sources in the text of Figure 4.2.

In yet other cases, students were given a diagramming tool with which they could make their own diagrams, like that in Figure 4.4, instead of the correct diagram. The tool appeared immediately below the text.

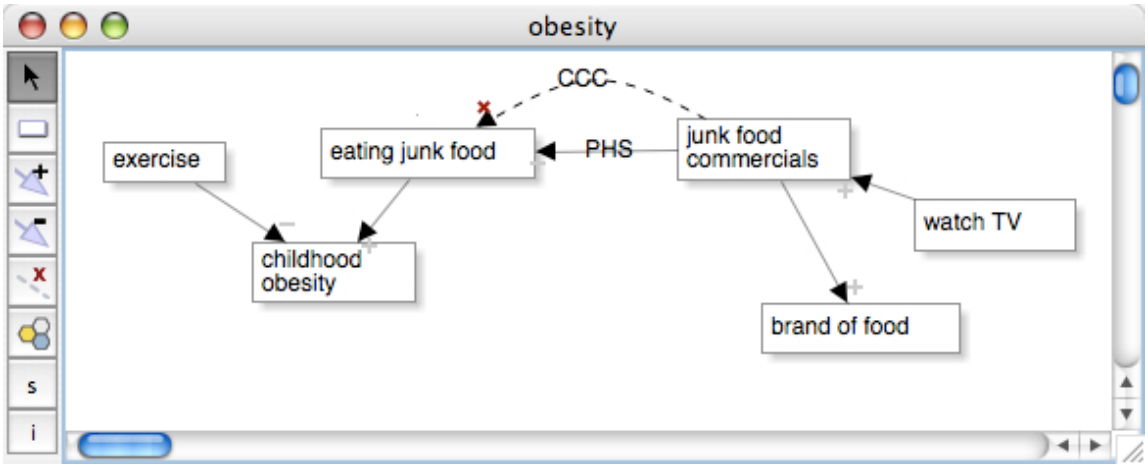


Figure 4.4. The iLogos tool with which students could construct their own diagrams.

Below the text, diagram, or tool, students were presented with 10 multiple choice questions in a randomized order (Figure 4.5).

Chain questions in which there is a causal chain from the first variable to the second according to the sources, and for which the correct answer is *yes*.

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?

Conflict questions in which the sources disagree about the causal path, and for which the correct answer is *inconclusive*.

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?

No path questions in which there is no causal path so the correct answer is *no*.

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?

Common cause questions in which a third variable causes the variables in the question and for which the correct answer is *no.*

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?

Common effect questions in which a third variable is caused by the variables in the question and for which the correct answer is *no.*

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?

Figure 4.5. Multiple choice questions asked after each test.

Students also received a set of training exercises telling them how to use the texts to answer the questions. For students who were given diagrams or tools, the procedure they were taught incorporated the rules for interpreting the diagrams (Figure 4.6). For example, 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 (Figure 4.6, 1). Which results in the diagram shown in Figure 4.6, 2. Then the student looks for a path between commercials and obesity, and finding no link (Figure 4.6, 3), answers *no*.

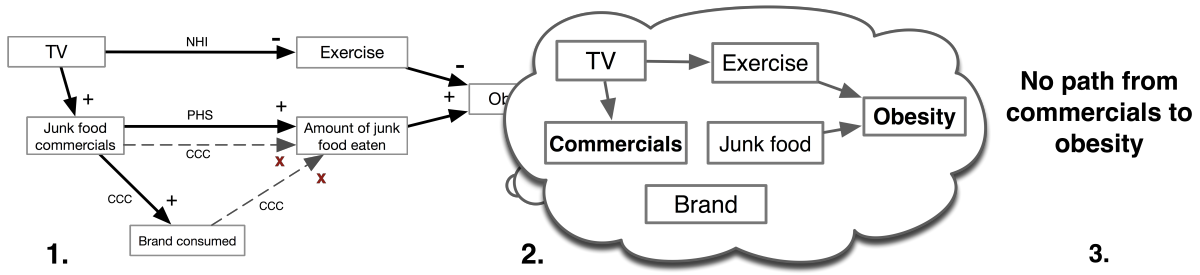


Figure 4.6. General procedure for using a diagram to answer a test question.

More precisely, diagram students were taught the following procedure:

- Read the question to identify the policy intervention variable and outcome variable.
- Read the question to determine which sources are relevant.
- Visually search the diagram to find paths from the intervention to the outcome using only arrows labeled with the credible sources and the unlabeled arrows representing common knowledge (while ignoring arrows labeled with other sources).
- Decide using the found paths where there is a path, no path, or conflicting path.
- Answer *yes* if there is a path, *no* if there is no path, or *inconclusive* if there are multiple contradictory paths.

Unlike the task in Chapter 3, the task in this study did not ask students to search for or evaluate information, because causal diagrams should not affect those stages of reasoning. Instead, the task focused on comprehension, synthesis, and decision through diagramming / non-diagramming paths (see Figure 2.2, in Chapter 2). 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 the diagrams should have the greatest influence on reasoning. If diagrams indeed prove beneficial, then future work will investigate more complicated, ill-defined tasks.

Procedure

Students were randomly assigned to either the text, diagram, or tool groups, then completed a pretest given in text, followed by a short training, then a performance test with text, text plus a correct diagram, or text plus a diagram tool, and finally a learning test with text only (Figure 4.7).

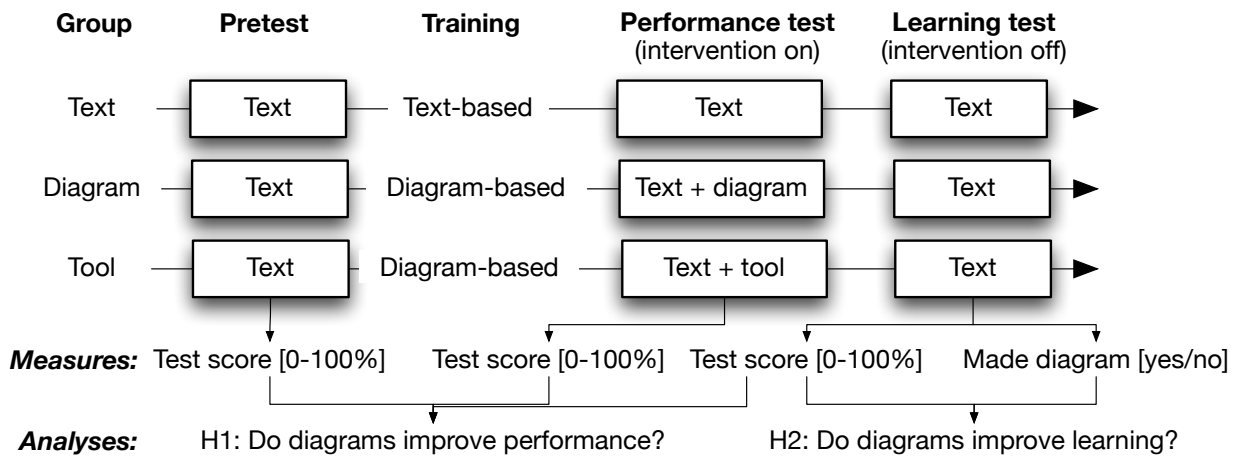


Figure 4.7. Experimental procedure, measures, and analyses.

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. The text was similar to that described in Figure 4.2. Students answered 10 questions about how intervening on one variable would affect another variable, according to different sets of sources, similar to those in Figure 4.5. The causal model in the pretest had a structure identical to that in the performance test and learning test, as well as questions on the same causal relations.

Students causal reasoning training consisted of five brief exercises. The purpose of the training was to introduce students to the concept of causal claims, and how to make inferences from multiple causal claims about the effect of a policy intervention on an outcome. For the text students, the training taught text-based procedures. For the diagram and tool students, the training taught diagram-based procedures.

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, e.g., "According to the NHI does smoking increase your chances of getting lung cancer?" Students received feedback with explanations immediately after each answer.

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, relevant claims were 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.

The **performance test** consisted of a 223 word text on junk food advertising and childhood obesity (Figure 4.2) 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 with which they could construct their own diagram. Students were asked the 10 multiple-choice questions in Figure 4.5. The purpose of the performance test was to test whether diagrams (or diagram tools) would help students to solve policy problems better than text alone.

The **learning test** consisted of a 201 word text on the deterrent effect of three strikes laws on crime that had the same causal structure as the pretest and performance test. 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 performance test. The purpose of the learning test was to see if there was a benefit to practice using the diagram, or if diagram tools helped students acquire skills that could be used even when diagrams were not provided.

Results

Effects of diagrams.

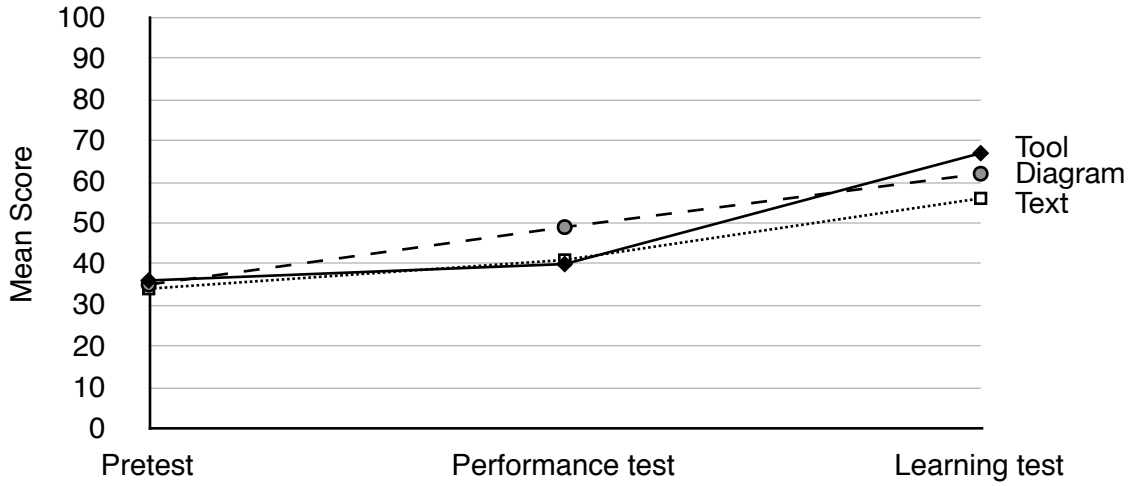


Figure 4.8. Test scores for text, diagram, and tool students.

The pretest, performance, and learning test scores of each group are shown in Figure 4.8. 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 **performance** test 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). Performance test scores were regressed on condition, time on training, and time on performance test. These three predictors accounted for 30% of the variance in performance test scores which was highly significant, $F(8,54)=4.30$, $p < .0005$. Both being given a correct diagram ($b=41$, $p < .04$) and spending a longer time on training ($b=3.2$, $p < .05$) significantly increased performance test scores. There were also two interactions. For students who had been given a correct diagram, there was an increase in performance test scores among those who spent a shorter time on training ($b=-4.8$, $p < .01$) and among those who spent a longer time on the performance test ($b=6.9$, $p < .03$).

Despite the superior performance of the diagram students on the performance test, on the **learning 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 learning test scores on condition and performance test scores 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 learning test scores than students in the text condition ($b=12$, $p < .03$), and students who had higher performance test scores also had significantly higher learning test scores ($b=5.0$, $p < .0000002$). Learning test scores of diagram and text 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 performance test.

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 learning conditional on making a diagram.

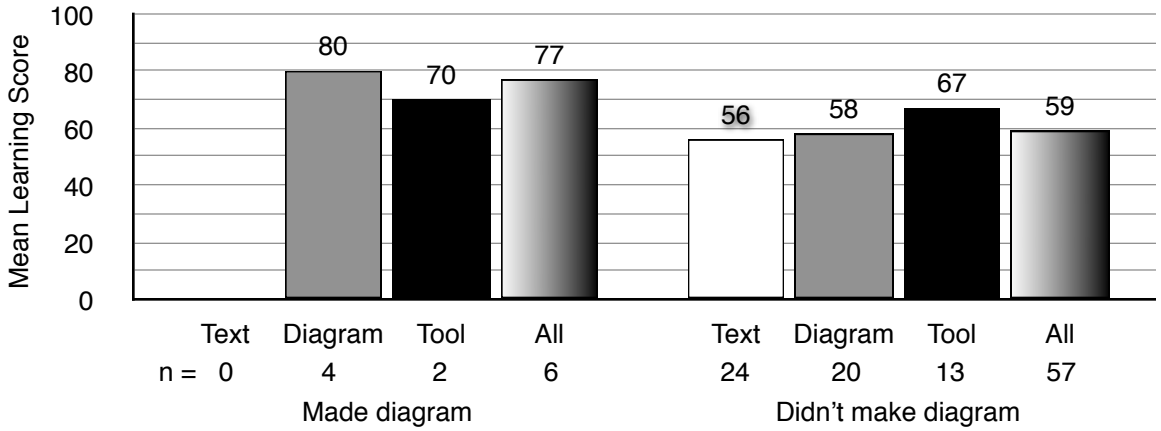


Figure 4.9. Learning test scores for students who at their own initiative made diagrams on the learning tests (left), and for those who did not make a diagram during the learning test.

To better understand the tool group's learning gains, I looked separately at students who made or did not make a diagram on scratch paper during the learning test (in Figure 4.9, I compared *all* students who *made diagrams* to *all* students who *did not make diagrams*). The six students who made diagrams on scratch paper had higher learning 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 learning scores (b=17, p<0.04), accounting for 5% of the variance, $F(1,61)=4.28$, $p < .04$. The higher learning test scores of students who made diagrams suggest either that diagrams are useful or a selection effect where the "good" students made diagrams.

The learning scores of the six students who made diagrams on scratch paper during the learning test (only two of which are from the tool group) cannot account for the higher learning test scores of the tool group as a whole. The tool group had higher learning 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 learning 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 learning test scores of students who did not make diagrams on condition and time on the learning 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 learning test scores ($b=14$, $p < 0.02$) than students in the text condition, and students who spent longer on the learning test ($b=5.3$, $p < 0.0004$) also had significantly higher learning scores. The results suggest that, when

diagrams were unavailable, having practiced constructing diagrams (on the performance test) led to higher scores on the learning test even if one did not make a diagram.

Time.

There were no significant differences in time between groups on the pretest, performance test or learning 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 performance test suggest that causal diagrams do indeed provide a good representational system for deliberation. Furthermore, even if students cannot, or will not, make diagrams on the learning 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; all experts declined payment.

Procedure

The procedure was identical to the procedure described in the previous 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, I 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 them.

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 five test questions, after which Expert 1 ended the experiment stating, "My brain is fried." While Expert 1's performance seems abysmal, recall that in this condition I have prevented him from using the causal diagramming tools that he was accustomed to

using. Unlike Novice 1, Expert 1 realized the futility of completing the task without a diagram. This performance underscores the difficulty of reasoning about even simple causal systems using text alone.

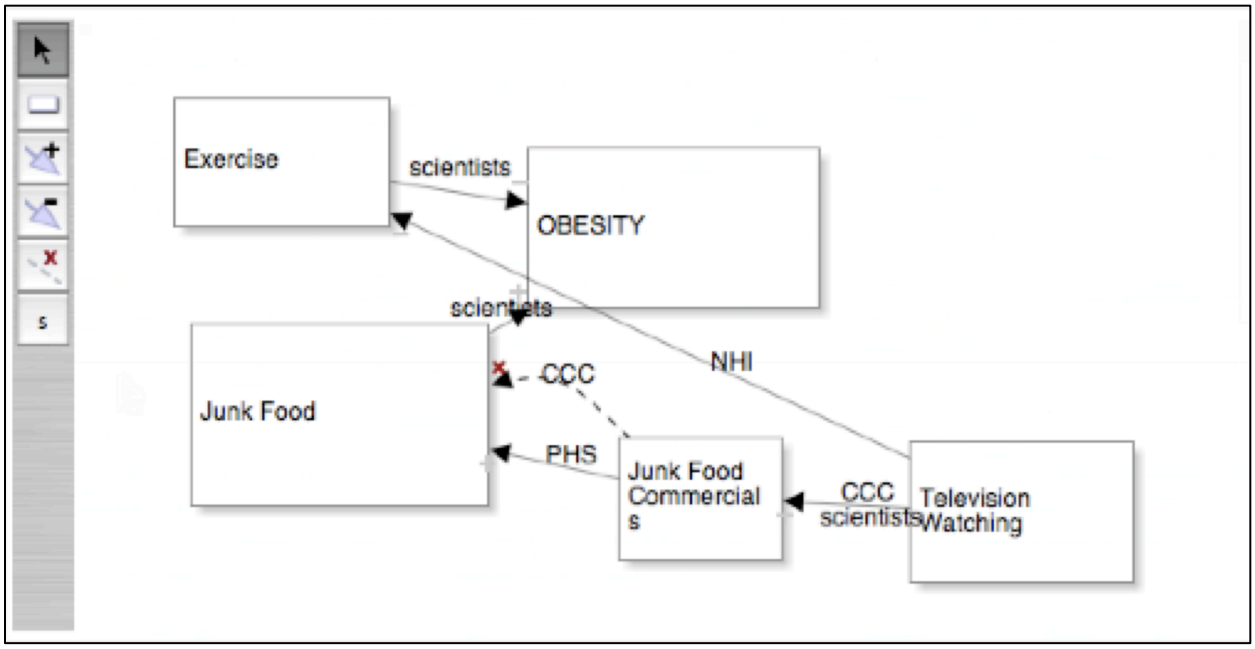


Figure 4.5. Novice 2's diagram.

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% on the performance test, performing no better than the text group. Given the poor performance of the diagram construction group in the previous study, I expected Novice 2 to have difficulty with diagram construction.

In fact, both Novice 2 and Expert 2 made better diagrams than those observed in the previous study. Figure 4.5 shows Novice 2's diagram. Compared to the correct diagram in Figure 4.3, Novice 2's diagram omits the *brand* variable, and mislabels some of the arrows, e.g., the arrows to obesity, and the arrows from TV to commercials should be unlabeled (representing common knowledge). Nevertheless, the diagram is a large improvement over the diagrams 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, Novice 2 might answer every question on the performance test incorrectly if he were to properly interpret his diagram. While both their diagrams contained errors, Expert 2's diagram (assuming it was used correctly to answer the test questions) would have lead to the correct answer on 100% of the questions, whereas Novice 2's diagram would have lead to the correct answer on 20% of the questions.

Interpreting diagrams (Novice 3 & 4, Expert 3).

The errors in the diagram condition were categorized in a bottom-up manner from observations of the protocol. These errors represent learning challenges (Figure 2.1) corresponding to the *decision*

(via interpretation) step (Figure 2.2). This analysis showed that participants' background knowledge and beliefs often interfered with their interpretation of the diagram (Table 4.1).

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: *According to the PHS, will making kids exercise more reduce the number of junk food commercials they watch?* Expert 3 correctly interpreted the diagram (Figure 4.3), 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.

While this answer would be an error by the grading criteria of the study, Expert 3's behavior could be considered 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, which is arguing for limits on junk food advertising, would accept the NHI's claim that TV affects exercise (*recent studies by scientists at the National Health Institute have shown that watching more TV does cause people to exercise less*):

> 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 as errors in the *construction* step of deliberation (Figure 2.2), rather than *decision (via 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. These errors also demonstrate how tightly intertwined construction and interpretation are, in the sense that as the expert is interpreting the diagram, he seems to be mentally reconstructing the diagram with his background knowledge. It may be that novices are to preoccupied with understanding the diagram syntax to dynamically critique it in this manner.

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. When answering question 7 (*According to the NHI, will making kids exercise more reduce the number of junk food commercials they watch?*) 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 [interpreting the NHI arrow saying that *TV* decreases *exercise* as meaning that *exercise* decreases *TV,* in which case *exercise* would decrease *TV* which would decrease *commercials*], 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 (in which she must, and does, recognize that lack of an arrow/path means the intervention does not affect the outcome). 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 4.1

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 the construction and decision via interpretation steps of deliberation (Figure 2.2) 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 the *decision (via interpretation)* step of deliberation, 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 this problem is warranted.

It may be possible to provide automated tutoring on synthesis and to detect normative and non-normative uses of background knowledge. This would require the tutor to: (a) provide a microworld in which students can collect new information, preventing any justification for speculation errors, (b) using confidence meters to detect and allow normative uses of background knowledge such as an override error, 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. Chapter 5 will show how such a deliberation tutor can be designed using the features described above.

CONCLUSION

The purpose of these studies 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, I 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, I 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 previous 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 did not test for the effects of construction practice on learning. This result shows that there are differential effects of receiving and constructing diagrams on performance and learning.

With respect to the third question, I 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.

These studies contribute 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, medicine, and more generally to domains in which people must construct representations of conflicting evidence from multiple sources such as argument, law, and intelligence.

Now, with a model of deliberation, with indications that we should focus on analysis, and with the knowledge that diagrams can help us overcome problems of synthesis, we can now turn to the design of a tutoring system for teaching policy reasoning.

5. An instructional system for teaching deliberation

Summary. The cognitive model and learning element studies described in the previous chapters provide a partial foundation for an instructional system that combines tutors and games to teach deliberation. However, knowledge of the learning elements does not fully determine the design of an instructional system. Furthermore, in the case of policy reasoning, ill-definition creates serious obstacles for the *inner-loop* of the tutor. There are difficulties in interpreting student input, the tutor's model may not always specify the correct solution, there are problems in assessing differences between the student's input and the tutor's model, and there are challenges in producing the feedback displayed to students. The first contribution of this chapter is to demonstrate how a combination of several tutoring strategies: (a) to reify, (b) to limit, (c) to tilt, (d) to use process constraints, and (e) to use student translation, can overcome problems of ill-definition. Overcoming these challenges of ill-definition along with a simple argument algorithm paves the way for the second contribution: a tutoring system that teaches deliberative argument. The third contribution is a pedagogical module that can dynamically switch between direct, cognitive, Socratic, stoic, and game-based modes of feedback. This pedagogical module makes it feasible to combine an intelligent tutor with a game-based inquiry environment, providing a platform for experimentation. These contributions advance cognitive tutoring across a number of ill-defined domains including policy reasoning (e.g., civics, political science, and public policy), domains that argue about causal systems (e.g., science, economics, and history), and more generally in domains that use diagrams to represent problems and organize evidence (e.g., argument mapping in philosophy and contextual modeling in HCI).

The previous chapters defined a cognitive model of deliberation, localized learning challenges in synthesis, showed how causal diagrams can help to overcome these learning challenges, and identified additional challenges associated with learning to use causal diagrams. This prior work lays a partial foundation for the design of a deliberation tutor.

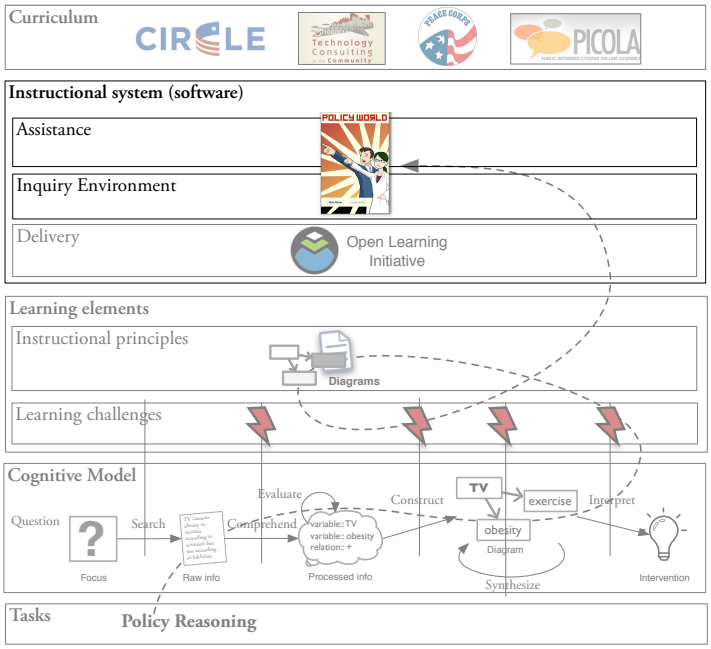


Figure 5.1. Chapter 5 asks: *How might we design an intelligent computer tutor to teach the skills of policy deliberation?*

However, even a fully specified account of the learning elements is not sufficient to determine the design of an instructional system, or even to guarantee that such a system can be designed. This chapter asks the fourth research question: *How might we design an intelligent computer tutor to teach the skills of policy deliberation?* (Figure 5.1).

This chapter describes the design of Policy World, an intelligent tutor embedded in an educational game for teaching deliberation. Deliberation is an ill-defined domain, so designing a deliberation tutor is much more difficult than designing a tutor for a well-defined domain like algebra.

The inner-loop of a tutor defines how the tutor provides assistance at the step level (as opposed to the problem level). Ill-definition creates challenges to providing assistance throughout the *inner-loop* of the tutor, in: (a) the student's **input**, (b) the tutor's **domain model**, (c) the tutor's **expert model**, (d) how the tutor **assesses** the student's action, and (e) the tutor's **feedback**. The student's input may create problems of ill-definition, for example, if the student inputs natural language that the tutor can't understand. The tutor's domain model may contain ill-defined information such as vague or contradictory terms. The tutor's expert model may be ill-defined, for example if the model does not define all correct solutions to a policy problem. Assessment may face obstacles of ill-definition, for example if the tutor cannot map between the student's input and the domain model. Even the tutor's feedback may be ill-defined, for example if the feedback is provided in natural language that must refer to terms input by the student which the tutor does not understand. Any ill-definition in upstream steps of the inner-loop creates difficulties for the downstream steps (e.g., if the tutor can't understand the student's input, then this creates problems for assessment and feedback). This chapter will describe how challenges of ill-definition arise in each step of Policy World's inner-loop, and how these challenges are addressed.

Assuming that these challenges can be overcome, then it may be possible for Policy World to provide tutoring on argumentation. While intelligent systems have not been able to tutor general argumentation, it may be possible to provide feedback if we restrict debate to causal arguments, presuming the initial obstacles of ill-definition can be overcome.

Finally, Policy World attempts to combine a tutoring system with a game environment, each of which suggest different approaches to assistance that cannot all be provided by the traditional cognitive tutoring architecture. Assuming that we can overcome the challenges of ill-definition and create an inquiry environment that can argue with students, we will then need a more flexible pedagogical architecture that can accommodate both tutoring and game-based approaches to assistance.

General strategies for addressing ill-definition

To address the problems of ill-definition, Policy World uses several strategies: (a) **reifying** the task in the interface, (b) **limiting** the task, (c) **tilting** the model away from ambiguous cases, (d) using **process constraints**, (e) relying on the **student to translate** between the student's and tutor's representations, and (f) **hedging**. *Reifying the task* means breaking the task into smaller steps according to the cognitive model and creating user-interface elements used by the student to perform each step. Breaking the task into smaller pieces requires adding definition to the task, or at least isolating the well-defined steps of the task, while the user-interface elements translate the student's work into a machine readable form. *Limiting* the task just avoids the difficulty, but can only be used

when a particular challenge arises in a problem-solving step that is not relevant to the learning objectives of the tutor. Limits can be placed on what the student is allowed to input, what the expert model allows the system to tutor, or the information in the domain model. *Tilting* means using problems that avoid some ambiguity, for example, providing evidence that that favors one possible solution more than another (without altering the nature of the skills practiced). *Process constraints* specify characteristics that a certain step should satisfy, but not the exact action. A process constraint is neither weak-adherence to a model nor a solution constraint as described in Lynch, Ashely, Aleven & Pinkwart (2006), but rather a midpoint between the two. The tutor can offload some of the burden of understanding ill-defined information by relying on *student translation* to convert a student representation into a representation that the machine understands, for example, by asking the student to describe their representation in different terms. Finally, the tutor can *hedge* its feedback by warning the student that there may be something wrong as opposed to providing an error message.

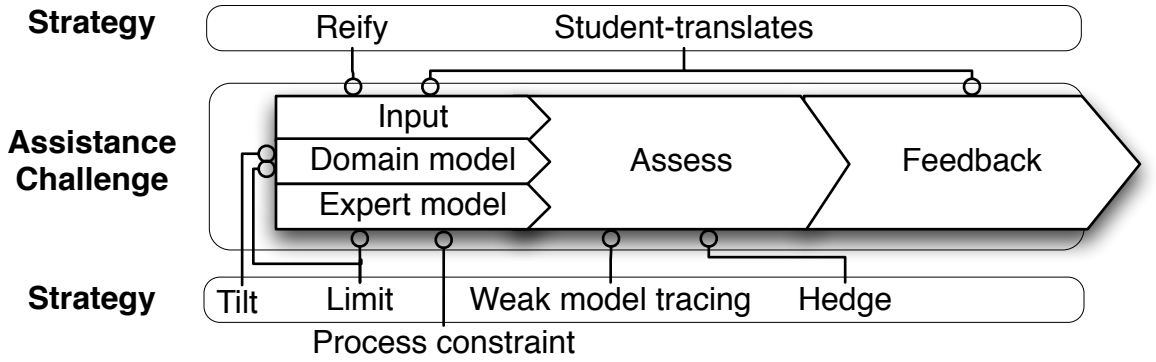


Figure 5.2. Challenges to providing assistance created by ill-definition (in input, domain model, expert model, assessment, and feedback) and intelligent tutoring strategies for addressing the challenges (reify, student-translates, tilt, limit, process constraints, weak model tracing, and hedging).

Figure 5.2 shows which challenges are addressed by which strategies. Each challenge and strategy will be described concretely later in the chapter.

General Policy World architecture

In order to explain how Policy World overcomes the challenges of ill-definition, we must first describe some of its components. Here it may be useful to compare and contrast it with its better-known cousins: the entertainment game Phoenix Wright, and the intelligent tutoring system Cognitive Tutor Algebra.

The inquiry environment

A game-based environment

Much of Policy World's design focuses on the inquiry environment (user interface). This is for two reasons. First, the user-interface is perhaps the most important tool for breaking problem solving into smaller, more discrete, more machine-readable steps necessary to add definition. Second, anecdotal feedback from students in the studies described in Chapter 3 and 4 suggested that they were more engaged in the policy task when it included even a minimal fantasy context (Chapter 3). Thus, Policy World employs a game-based inquiry environment.

Figure 5.3. Screenshots from *Phoenix Wright: Ace Attorney* showing search for evidence, interrogation of a witness, and debate (Capcom, 2005).

The inquiry environment of Policy World borrows heavily from *Phoenix Wright: Ace Attorney*, a popular, single-player, "visual-novel" adventure game where players assume the role of an attorney (Figure 5.3). In Phoenix Wright, players search for evidence by clicking on pictures of crime scenes (Figure 5.3, left) and by interviewing witnesses (Figure 5.3, center). After searching for evidence, players defend their client in a courtroom trial. The courtroom trial primarily consists of making objections to particular claims made by witnesses. When the player makes an objection, he must then support the objection by producing a piece of evidence, such as a murder weapon, that somehow contradicts the claim. The player does not have to explicitly state how the evidence contradicts the claim, but they do need to know which claim to object to and which piece of evidence to provide.

The cover story of Policy World mirrors that of Phoenix Wright. In both Phoenix Wright and Policy World, half the interface includes a static background image with a character delivering some dialogue. The interface also includes controls for the discrete set options the player can perform at that particular moment such as a button to “search for evidence” or a list of possible responses to computer characters such “agree” or “disagree”. In Policy World, the student plays a young policy analyst who must persuade a senator to adopt his policy positions. Instead of competing against an unscrupulous lawyer, students in Policy World compete against an unscrupulous lobbyist (played by the computer). The underlying narrative themes are also similar to Phoenix Wright: the role of the student’s character is to defend justice (or in this case social development) while progressing from novice to expert.

Unlike Phoenix Wright, Policy World is designed for education. Policy World is intended to help students learn the skills of deliberation, and the system is designed both to teach and to assess these skills. Like Phoenix Wright, each level of Policy World consists of a specific case in which the player must search for evidence and make arguments. In Policy World, students spend much more time analyzing evidence (in Phoenix Wright players do not analyze evidence) and constructing much more complicated arguments. Each level of Policy World consists of a case, such as “Should we

decrease junk food advertising on children's television?" The first level contains a short tutorial and (unbeknownst to the student) a pretest. Levels 2-4 consist of training levels with cases of increasing complexity. The final levels of Policy World contain two posttests. The pretest and posttests offer less assistance than the training levels which the story explains away as the consequence of external events, e.g., a mentor character becoming unavailable.

Intelligent debaters

A fantasy context may promote engagement, but it does not necessarily give students a *reason* to engage in search and analysis of evidence. Fortunately, having students debate an opponent gives them a reason to search and analyze evidence. In addition, debate requires students to explain the reasoning behind their policy recommendations and to cite evidence to support their positions. Situating the search and analysis tasks as preparation for debate makes these tasks simultaneously more authentic and more game-like. So adding debate may help us both increase motivation *and* learning.

Creating an intelligent tutoring system that can debate the student is not straightforward. As described in Chapter 2, other games have not fully succeeded in creating authentic debating tasks. Furthermore, policy reasoning is an ill-defined domain which raises a host of additional challenges. For example, if students are creating their own diagrammatic representations with their own terms as in Chapter 4, then the tutor must be able translate between the student's representation and the knowledge encoded in the tutor's domain model.

To debate the student, Policy World uses a set of *intelligent debaters* that argue with the student about policy. Students are asked to make policy recommendations, to provide causal mechanisms that explain how their recommendations impact the relevant outcomes, and to cite evidence in support of their positions. Students argue with the debater using the terms of the expert model to avoid ill-definition caused by providing feedback in natural language. However the system provides assistance using the student's terms. This requires the tutor to map between the student's representation and the representation in the tutor's domain model. I will describe the details of this approach later in the chapter.

Assistance

As in many argumentation systems, the tutoring system in Policy World consists of a diagnosis module and a pedagogical module (Scheuer, McLaren, & Pinkwart, 2010). The diagnosis module determines whether the student's actions deviate from a normative problem-solving process, which corresponds to the expert model of a cognitive tutor (VanLehn, 2006). The pedagogical module responds to the student's errors by providing assistance, primarily through Socratic questioning. Policy World's diagnosis and pedagogical modules were designed specifically to overcome challenges of ill-definition not faced by intelligent tutors in well-defined domains like Algebra.

The diagnosis module

The diagnosis module of Policy World differs from that of a cognitive tutor in several ways: (a) it does not use a Rete-based production system engine, although knowledge is represented in a rule-based form, (b) the student is assessed using *process constraints* rather than with model-tracing or constrain-based tutoring, and (c) rules represent the tutor's knowledge, not the experts. These

differences arise from from both the practical and theoretical challenges of tutoring in an ill-defined domain.

Because the inquiry environment presented the greatest challenge, Actionscript/Flex was chosen as the programming language, because its user interface toolkits and cross-platform player provided certain advantages. Unfortunately, Actionscript lacks an established Rete-based production system engine, so Policy World tutor does not use a list processing production system. Nevertheless, the diagnosis and pedagogical behavior of the Policy World tutor is represented in a rule-like form similar to that of a cognitive tutor (Koedinger, Anderson, Hadley & Mark, 1997). The rule objects in the diagnosis module each possess a "matches" method (e.g., left hand side of a production rule) that determines whether the rule applies to the current problem state, and a "fires" method (e.g., right hand side of a production rule) that defines the response of the system to that state. These rules also add and remove goal objects to trace the problem solving process.

A second difference from cognitive tutors is in how the Policy World tutor traces student actions. With respect to the expert model, it is necessary to deviate from the model-tracing approach, because in game-based environments, the student is allowed to deviate arbitrarily far from the correct solution path. This would make it difficult to write buggy productions for every incorrect path. Furthermore, at some points in the deliberation task, it is easier to assess whether an action is incorrect rather than whether it matches one of the possible correct solutions. By using constraints that take into account the subgoals in working memory, the tutor can allow the student to deviate from the correct solution path while still providing strategy-based feedback. In this case, the "matches" method of a rule object acts like a constraint on a step-level action, rather than a constraint on the solution as in a constraint-based tutor (Mitrovic, 2001). Unlike a solution constraint, the constraints in Policy World are relative to what the student should do given what he has already done. For example, if the student's diagram were to incorrectly claim that exercise increases obesity, then a diagram interpretation constraint would require the student to propose decreasing exercise in order to decrease obesity. These constraints are also more general than buggy productions. A better way to think of these constraints is as *process constraints*, i.e., each time the student does something new, the tutor checks to see if the student has violated any of the relevant process constraints. These constraints perform a model-tracing function and are thus located within the diagnosis module (i.e., the expert model).

The rules in Policy World also deviate slightly from those in a cognitive tutor. Rather than represent the knowledge of an expert problem solver (as in a cognitive tutor), the rules represent the knowledge of a tutor. For instance, the expert model in a cognitive tutor might contain a rule like: "IF there is a goal to add 2 and 2, THEN type 4." When the student does not type "4", or respond in a way consistent with the rule, the pedagogical response of the cognitive tutor is to inform the student of the error. A Socratic tutoring system like WHY (Collins, 1977) represents the rule from the tutor's perspective. For example, from a tutoring perspective, the corresponding rule might be: "IF there is a goal to add 2 and 2 AND the student does not type 4, THEN ask the student to identify the first addend." In other words, the rules specify what pedagogical actions the tutor should take, rather than what action an expert problem solver should take.

The rule object methods representing the "right hand side" create a list of error objects describing what the student has done wrong, rather than the particular action the tutor should take. The

pedagogical model is then responsible for deciding which errors to respond to (if at all) and what form the tutor's response should take.

The pedagogical module

Like a cognitive tutor, Policy World is designed to provide context-sensitive assistance based on the current goals of the problem solver, but it achieves this in a slightly different way. In a cognitive tutor, the diagnosis module (the expert model) models all possible correct actions. The cognitive tutor's pedagogical module simply informs the student when their action does not match one of the expert model's possible actions. The cognitive tutor author can add additional assistance by writing a buggy production that specifies a set of hint messages that the student can access by voluntarily clicking on the *hint button*. In contrast to a cognitive tutor, the pedagogical module in Policy World was designed to dynamically employ different pedagogical strategies in order to make the system more game-like as well as to test a range of hypotheses about how different levels of tutoring, varying in immediacy and directness, differentially affect learning and motivation.

To provide tutoring, the pedagogical module monitors the diagnosis module waiting for a constraint to be violated, i.e., for the student to make an error. When a constraint has been violated, rules in the pedagogical module determine which *question* (or set of *questions*) should be asked. Each question object can be thought of as a mini-production system consisting of an initial prompt (either a question or feedback message) and 2-6 productions for responding to all possible student inputs. For example, a simple feedback or error message might include a prompt such as, *Good job!* and only allow the student to respond by acknowledging the message. A Socratic question might use a prompt like, *What should we do to decrease childhood obesity?*, and allow the student to pick a policy recommendation from a list of possible options. After the student responds, the question determines whether or not the student's answer is correct, produces an appropriate feedback message if needed, and either exits, re-asks the initial prompt, or pushes additional questions onto the question stack. The tutor then asks whichever question is next on the stack, whether that be the current question, the next question, or a new sub-question. If the question object asks a sub-question, then the student must answer the sub-question correctly before the original question is re-asked. Figure 5.13, 5.17, and Table 5.2 will illustrate this process later in the chapter.

This pedagogical module architecture allows for several modes of tutoring simply by creating different question objects. These tutoring modes include:

- *Direct* tutoring, where the tutor tells the student a specific piece of information or specific command. Direct tutoring is used in Policy World when teaching students how to use the interface at the beginning of the game. During direct tutoring, the tutor pushes a list of questions into the stack, where each question must be answered before proceeding. The question prompts in direct tutoring give the student a command. After the student has successfully responded to the command, the question is removed from the stack. Because direct tutoring questions are designed to deliver tutorial-like instruction (rather than assist in problem solving), these questions do not typically add additional questions to the stack.
- *Cognitive* tutoring, where the tutor provides error flagging immediately after each step along with additional knowledge-based teaching feedback and hints. During cognitive tutoring, the tutor responds immediately to the violation of any constraint by pushing a question onto the stack. The question prompt will include immediate knowledge-based feedback.

- *Socratic* tutoring, where the student can attempt a combination of steps simultaneously, and the tutor responds to errors by dynamically scaffolding the sub-steps. When the student answers a question incorrectly, the tutor responds with a sequence of sub-questions. If one of these sub - questions is answered incorrectly, then the tutor asks another set of sub-questions, and so on, until the tutor has drilled down to the particular step causing the error. This approach creates a sort of on-demand scaffolding without slowing down students who have become proficient.
- *Stoic* tutoring, where the tutor intervenes only after critical errors. In this mode, the student may wander arbitrarily far from a correct solution path before the tutor intervenes. In this mode the pedagogical module responds only to violations of high-priority constraints, ignoring low-priority constraints.
- *Game* (no) tutoring, where the tutor does not intervene, and the student receives only the feedback provided by the inquiry environment. In Policy World, the tutor is not the only means of providing assistance. The game itself provides some situational feedback. For example, at certain points of the game the student debates a computer opponent who may critique some of the students argument moves. This feedback is provided implicitly by the inquiry environment as opposed to explicitly by the tutor. The game scoreboard also provides minimal error flagging even though it is not part of the tutor. During game tutoring, the student receives only the situational feedback of the game and the minimal error flagging of the scoreboard. In other modes of tutoring, such as cognitive tutoring, the tutor can decide whether the student should also receive the situational and minimal feedback in addition to the tutor's feedback. The tutor can also decide to take control of the inquiry environment and provide only its own feedback.

Unlike a traditional cognitive tutor, a compelling game-like tutoring system requires each of these modes of assistance, and the question-based pedagogical module allows us to move seamlessly between these modes using a common architecture.

Of course, the diagnosis and pedagogical modules of Policy World are only useful for a deliberation tutor if we can overcome the problems of ill-definition that arise in the domain of policy reasoning.

Policy World walkthrough

To demonstrate how Policy World overcomes the assistance challenges created by ill-definition by reifying, limiting, tilting, using process constraints, and using student translation, I will provide an example in the form of a walkthrough of the Policy World game. The description will focus on how the inquiry environment reifies the cognitive tasks specified by the deliberation framework, obstacles to providing assistance, and the strategies for overcoming these obstacles.

Prologue

Figure 5.4. The prologue of Policy World introduces the player to all of the game characters including her boss, the senator, her opponent the lobbyist, her mentor, and computer tutor (not shown).

In Policy World, students play the role of a policy analyst who must make evidence-based recommendations to the senator about policy topics like the 21 age drinking limit. At the beginning of the game (Figure 5.4), the student's boss welcomes her to her new job at a policy think tank and tells her that, in order to "save the country," she will need to make policy recommendations to the senator who requires evidence-based analyses in order to make decisions. The student is also warned that different opponents, including a lobbyist named Mr. Harding, will oppose her recommendations. The student is introduced to a mentor character who will lead her through several training problems and a computer tutor that will teach her how to analyze information. Early in the game, the mentor character solicits the student's pre-existing beliefs about policy topics. The game will then alter the sets of evidence available on each level so that the majority of the evidence contradicts the student's prior beliefs, as in Chapter 3.

After students have been introduced to the game, they complete a series of problems, some of which test their learning and others that provide instruction. Each problem/level consists of two phases: in the first phase, students search and analyze evidence; in the second phase, they debate a computer opponent.

Question

At the beginning of each problem, the student is given a policy question like, "Should we reduce the drinking age to 18?" This is provided to the student in the form of a policy brief (Figure 5.5). The brief specifies: (a) a *policy goal* which typically specifies the policy outcome that should be increased or decreased, (b) an *initial question* which may identify the possible interventions the student should consider, (c) *issues* which are variables that must be included in the student's final explanation of their policy recommendation, and (d) background material describing the basic ideas for students who are completely unfamiliar with the topic.

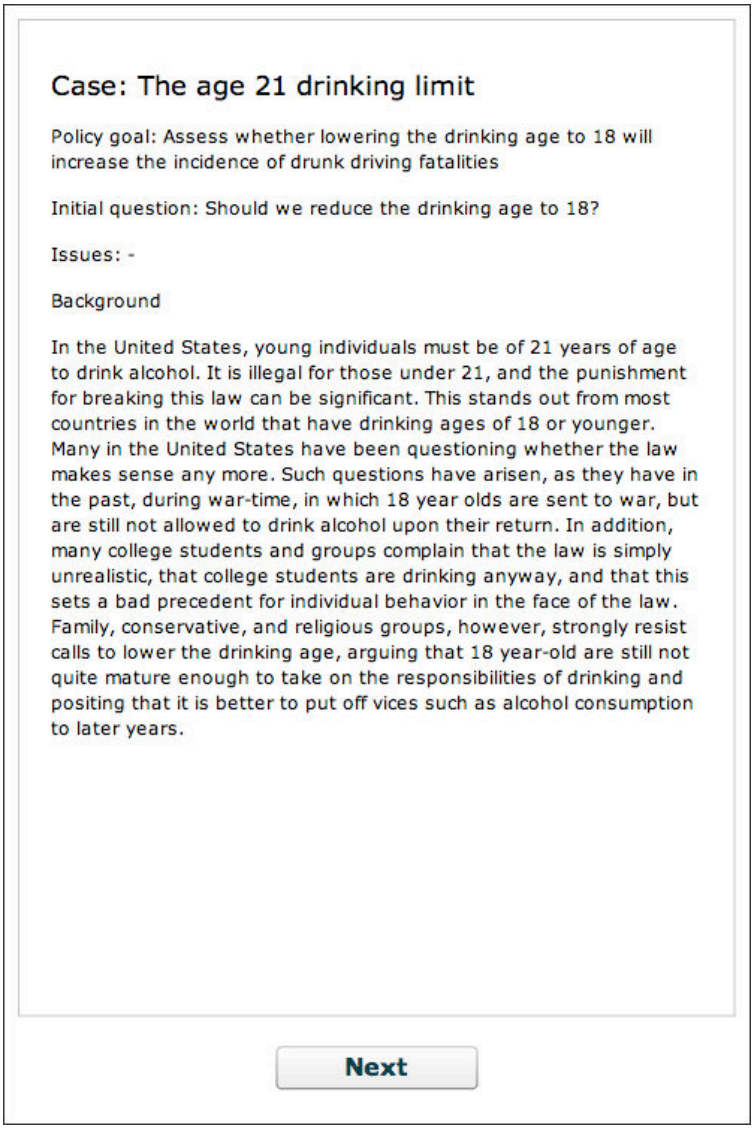


Figure 5.5. Policy brief for the drinking age problem.

The policy brief corresponds to the question/focus step in the deliberation framework described in Chapter 2 (Figure 2.2).

In this initial version of Policy World, the question for a given problem/level is fixed. Students cannot alter the focus of the problem, and at the end of the level they will be asked to debate the same policy question or one very similar to it. One could imagine a more realistic version of Policy World in which, after investigating the initial question, students can reframe the question or focus on a new, perhaps more fundamental policy question. For example, the student might begin with a question about cap and trade approaches to global warming, but determine that the issue cannot be resolved without first addressing campaign finance reform. In this version of Policy World, this can happen only to very a limited degree such as when the student discovers interventions that are not described in the policy brief. However, more significant reframing, such as convincing the senator to debate a different question, is not possible.

This assistance challenge created by the ill-definition in reframing concerns the tutor's *expert model*. Here the strategy is to *limit* what the student is allowed to do, i.e., the system does not allow student to substantially reframe the question. I will return to reframing in future work.

Search

After receiving the policy brief, the student must search for evidence using a (fake) Google interface (Figure 5.6, upper left). The student begins with a search term provided by the game such as "drinking age" that cannot be altered. After clicking "search", the student is presented with a list of search results such as the sites: *Why 21*, a pro, legal-21 site, *Choose Responsibility,* an anti, legal-21 site, and *The New York Times* (Figure 5.6, upper right). The student can visit each of the sites that appear in the Google results. Most of the websites have a homepage indicating the policy orientation of that site along with a list of reports, as in Chapter 3. The reports available to students are short, approximately 3-5 paragraphs, newspaper-like summaries ranging from interviews of policy advocates to summaries of empirical studies from the science section of a newspaper. Unlike the experiment in Chapter 3, most of the sites and reports in Policy World are adapted from real websites and real articles. Each report contains one or more causal claims, such as: *The study found that the recent spike in drunken driving, after years of declining fatalities was associated with increased binge drinking*. Each claim is associated with a source, such as: *Center for Disease Control*, and can be one of several evidence types, such as an experiment, observational study, case, or claim/belief.

The search task corresponds to the search step of the deliberation framework described in Chapter 2.

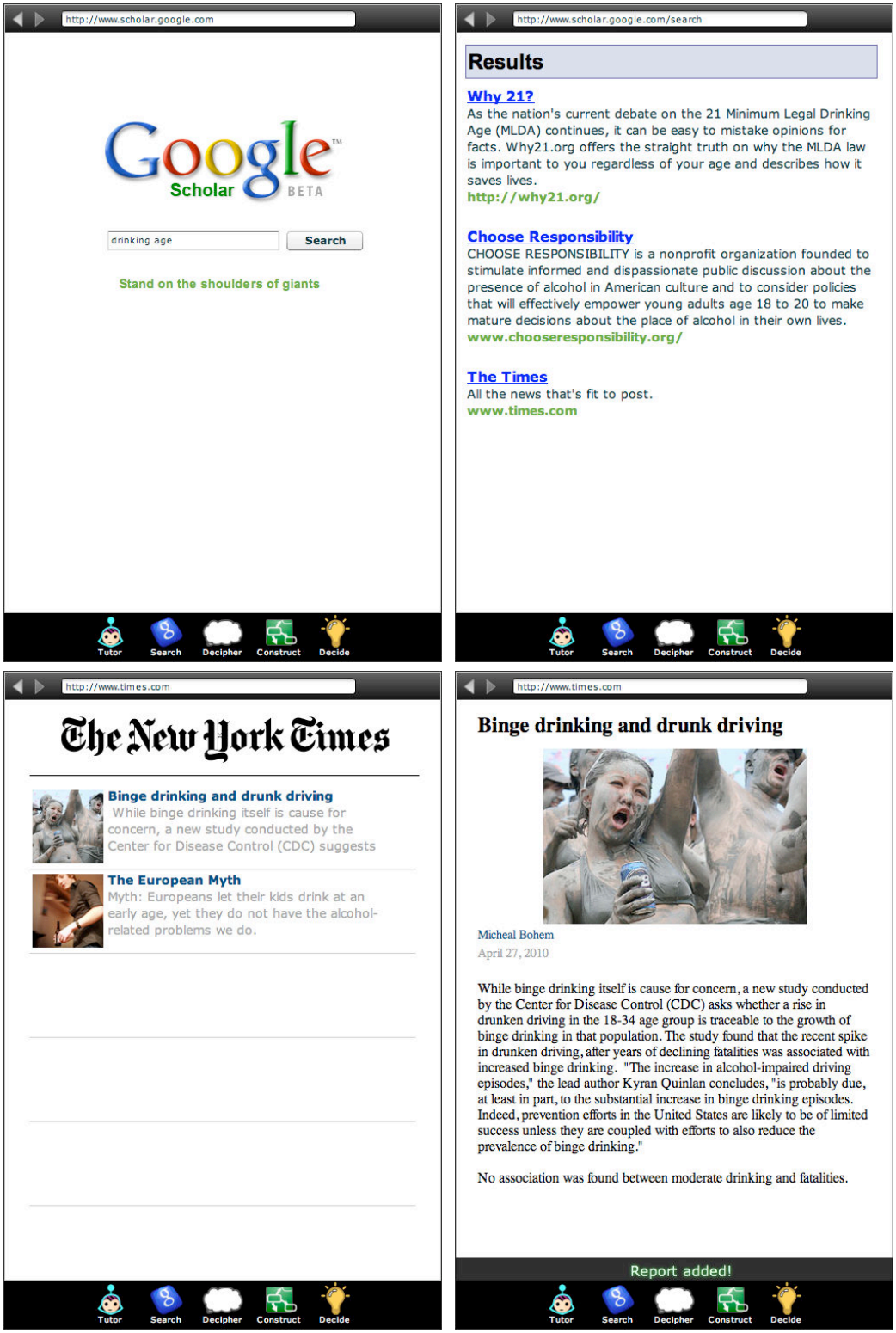


Figure 5.6. The fake Google interface used by the the student to search for information. Each site contains short newspaper-like summaries containing causal claims from different sources.

Several challenges arise in search, both in the tutor's *domain model* and its *expert model*. The first issue concerns the size of the search space in the domain model. In real policy problems, there is an

infinite amount of information, much of it which may be irrelevant to making policy decisions based on evidence, for example ad-hominem attacks on an opponent.

Chapters 3-4 argue for focusing instruction on analysis. While Taber and Lodge (2006) found that students were biased in their information search, tending to avoid sources that would contradict their beliefs, the study in Chapter 3 did not find the same bias using a similar task. As a result, Policy World attempts to minimize the amount of time the student spends on search and emphasizes analysis, unlike many of the science games built on virtual worlds. The study in Chapter 3 showed that with a simplified policy reasoning task, students could search adequately (in an unbiased manner) but could not analyze adequately (accurately synthesize evidence). So while the information that students can search for in Policy World is realistic, the amount of information to search for is greatly reduced. So the strategy for dealing with assistance challenge created by the ill-definition in search space size is to *limit* the information to a more tractable set. This a reasonable approach only because Policy World emphasizes instruction in analysis.

A second issue of ill-definition arises in the tutor's *expert model* and concerns the process of search, specifically, what stopping rule for when search is complete should the expert model teach? Ideally, one would like the student to perform a comprehensive, unbiased search for information, relative to the importance of the question and taking into account bounds on time and attention. Given the debate context, a rough approximation of the stopping rule might be: *Do I have enough information to beat my opponent?* Of course, that heuristic slightly reduces the problem of ill-definition to an estimation of what information one's opponent is likely to have.

The expert model of Policy World monitors the student's search and provides feedback such as: *You visited The Times, but you didn't look at all the relevant reports there*, when the student ends search before finding enough evidence to make a strong case during the debate. However, given the limited search space of Policy World's domain model, this tutor's feedback reduces to "search everything." This rule works in this case because the size of the search space in the domain model has already been severely limited, and would not work for a real policy problem. Again, the strategy of *limiting* the domain and expert model avoids the assistance challenge in search created by ill-definition, but is only allowable because search is not the focus of instruction.

Analysis

At any point before starting the debate, the student can choose to analyze one of the reports that she's found. To begin analysis, the student clicks on the *Decipher* button which presents the student with a list of reports (Figure 5.7, left). The student then clicks on a report to begin analyzing it. Analyzing a piece of evidence in Policy World requires four steps: *comprehension*, *evaluation*, *construction,* and *synthesis*, as described by the deliberation framework.

Comprehension

When the student chooses a report from the list of reports, she is presented with the full text of the report (Figure 5.7, right). The student then ***comprehends*** a causal claim by selecting the text of the claim in the report (Figure 5.7, right). For example, the student chooses the report: *Brief history of the drinking age,* and selects the text: *the 2006 Monitoring the Future observational study shows that 21 minimum drinking age laws decrease underage consumption of alcohol.* The texts in Policy World are

condensed versions of actual newspaper articles, investigative journalism reports, and editorials and finding the causal claims is by no means trivial (as will be seen from analysis of student errors in Chapter 6).

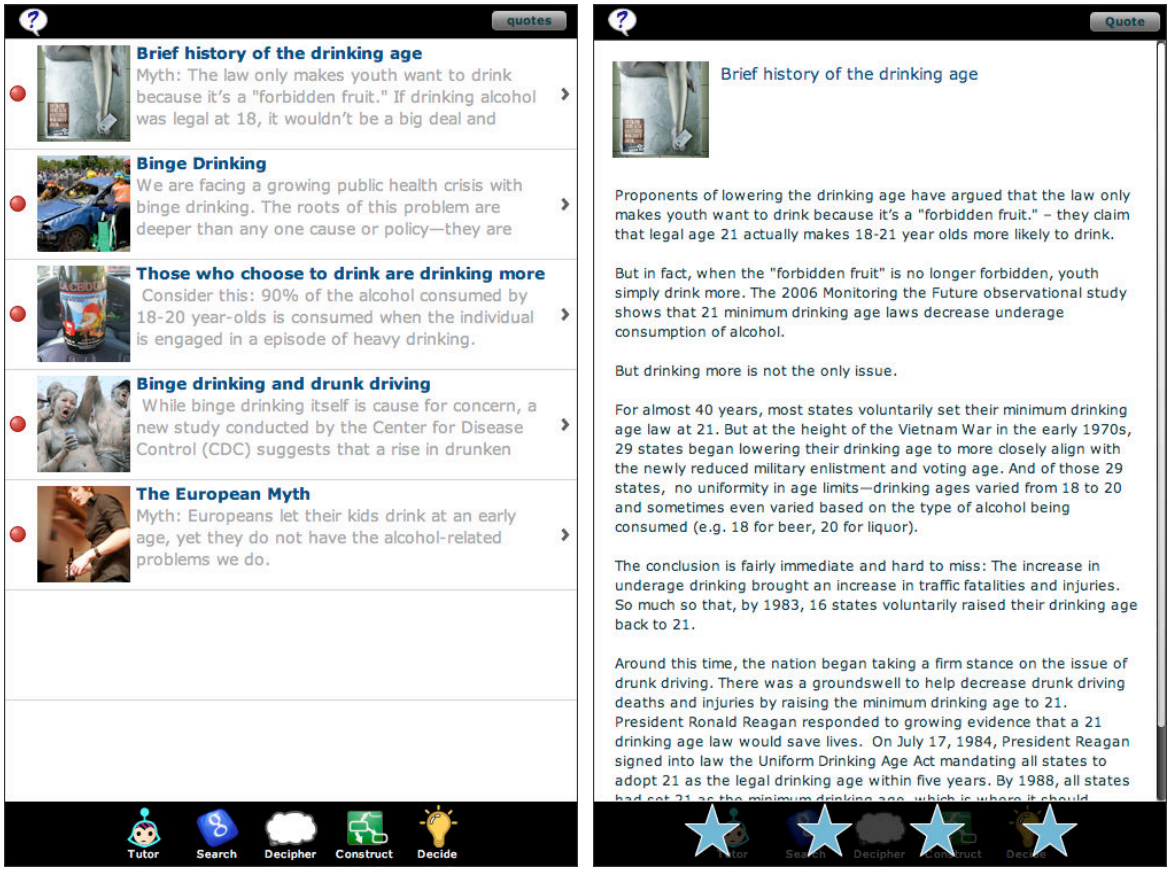


Figure 5.7. The interface for analyzing a causal claim in Policy World requires the student to choose a report (left) to select a causal claim in the text of the report (right).

Finding causal claims in text corresponds to the *comprehend* step in the deliberation framework (Figure 2.2).

In the comprehension step, the tutor's challenge is to determine whether or not the student's selection corresponds to a causal claim. This is primarily an issue for the tutor's domain model and in student input as well. In order to assess comprehension, each sentence of the reports in the domain model is encoded by an expert who identifies which sentences correspond to which causal claims. The expert also determines whether two different sentences are using different terms to refer to the same variable. Sentence boundaries are also coded. The expert model judges the student's selection to be correct if the selection: includes at least part of a causal claim, does not include multiple claims, and does include text from more than three sentences. This makes the the comprehension interface quite usable. Students aren't penalized for slightly sloppy selections that cross a sentence boundary or for selecting only part of a causal claim that spans multiple sentences. Students are also prevented from gaming the comprehension step by selecting huge blocks of text. To simplify tutoring, repeated, non-adjacent causal claims were removed from the text, and sentences that contained multiple causal claims were split.

Expert encoding is a necessary part of designing intelligent tutors and not unique to ill-defined domains. The ill-defined strategies include the slight cleaning of the text, which is a type of *limiting* strategy applied to the domain model, and the flexible sentence selection assessment, which is a kind of *weak model-tracing*.

Evaluation

After students have comprehended a causal claim, the next step is to ***evaluate*** the causal claim. To evaluate a claim, the student must use a combo box to specify the evidence type, e.g., an experiment, observational study, case, or claim. The student must also rate the subjective strength of the claim on a nine point scale labeled with the categories: *none*, *weakest*, *weak*, *decent*, *strong*, *strongest* (Figure 5.8). For example, if the student is evaluating the claim: *the 2006 Monitoring the Future observational study shows that 21 minimum drinking age laws decrease underage consumption of alcohol*, then the correct evidence type is *observational study*, and the strength rating should be on the high end, because observational studies are second only in strength to experiments, given the available evidence type categories.

Figure 5.8. The interface for evaluating a claim in Policy World requires the student to specify the evidence type of the causal claim and the strength of the causal claim.

The tutor's domain model encodes the evidence type of each claim, so it is easy for the tutor to provide feedback on evidence type. Providing feedback on evidence strength runs into challenges of ill-definition. There is no normative theory of how to quantitatively rate the strength of a piece of evidence. However, there are two heuristics that should not be violated, all other things being equal: (a) students should rate experiments as stronger than observational studies, as stronger than cases, as stronger than claims, and (b) students should not rate evidence as extremely strong when it is congruent with their beliefs and as extremely weak when it is incongruent with their beliefs. Of course in the real world, the problem of evaluation is much more complicated. For example, one might rate an empirical study by an interested, discredited source as weaker than a claim presented by a respected expert. Additional information about sample size and ecological validity might make an observational study more compelling than a small laboratory experiment. On the other hand, the texts from the *New York Times*, *PBS,* web-based advocacy sites, and other sources from which Policy World's domain model was constructed seldom contained the level of detail necessary for more sophisticated approaches to evidence evaluation. In other words these two simple heuristics for evaluation might not be completely unreasonable. Students employing these heuristics would certainly demonstrate better reasoning than that observed in Chapter 3.

The challenge posed by ill-definition in rating evidence strength lies in the expert model. Policy World dictates only that the student should rate evidence in an unbiassed manner and observe a particular ordering of evidence types. To implement these heuristics, Policy World uses *process constraints.* The diagnosis module contains a process constraint that runs whenever the student evaluates a claim and checks to make sure the student's strength rating does not violate either of the heuristics. It also checks that the strength rating will not create future violations (e.g., if the first causal claim is a *case*, the student cannot rate that evidence as strongest because there would be no valid strength rating for a later piece of evidence with the type *experiment*). As described earlier, this process constraint is not weak model-tracing, because the expert model does not contain a correct answer. The process constraint is also not a constraint on the solution as in a constraint-based tutor, because the process constraint concerns an action on a particular step which is sensitive to the student's current goals and previous actions. The process constraint thus allows strategy-based feedback, like a cognitive tutor, in an ill-defined domain.

Construction

After evaluating a causal claim, the third step of analysis is to ***construct*** a diagrammatic representation of the claim. During construction, the student uses boxes to represent the variables in the claim and arrows to represent the causal relations. For example, for the claim: *the 2006 Monitoring the Future observational study shows that 21 minimum drinking age laws decrease underage consumption of alcohol,* the student would create two boxes such as: *legal 21 drinking age,* and *drinking*, with an arrow from *legal 21 drinking age* to *drinking* labeled with a minus sign (Figure 5.9). After the student has constructed a diagrammatic representation of the claim, she links the arrow to the text quoted in the comprehension step. She must explicitly link the arrow to the text in order to avoid duplication, e.g., if she has already constructed an arrow for that causal relation. If the relationship has already been identified, she can skip the creation of additional boxes and arrows. In that case, she simply selects the relevant arrow and links it to the evidence using the orange button on the diagramming toolbar in Figure 5.9. Multiple pieces of evidence can be linked to a single arrow. During the debate, the student can click on any arrow in the diagram and see which pieces of evidence are linked to the arrow.

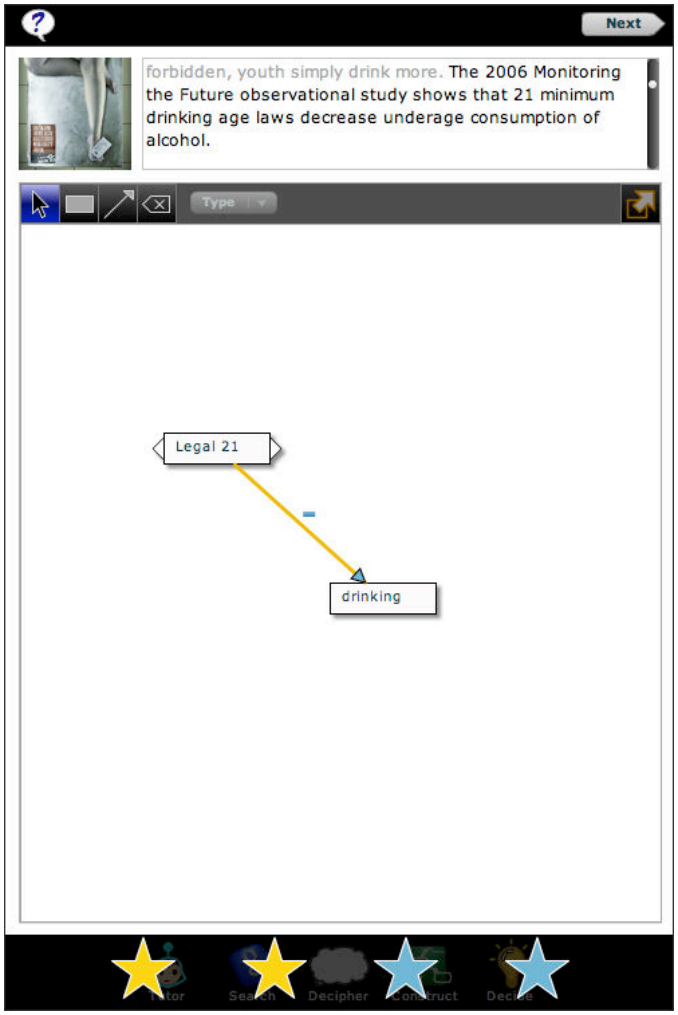


Figure 5.9. The interface for constructing a diagrammatic representation of a causal claim in Policy World. The student represents variables with boxes and causal relations with arrows. The student must link the causal arrow to the evidence being diagrammed.

While it is easy for a human tutor to provide feedback on diagram construction, it is more difficult for a computer tutor that lacks any natural language understanding. From the computer's perspective, it is like being an English-speaking tutor trying to provide feedback on a diagram written in Vietnamese. The problem of ill-definition here is in the student's input.

To provide feedback on diagram construction, the tutor first ensures that the student has linked the causal arrow in their diagram to the evidence they've cited from the report. This allows the tutor to translate from the student's representation to the causal claims encoded in the tutor's domain model. For example, if the student is analyzing the causal claim, *the 2006 Monitoring the Future observational study shows that 21 minimum drinking age laws decrease underage consumption of alcohol* and creates the two boxes: *legal 21* and *drinking,* the tutor cannot determine which box refers to which variable. When the student creates an arrow from the *legal 21* box to the *drinking* box, and links it to the causal claim, then the tutor infers that *legal 21* refers to *21 minimum drinking age laws* and that *drinking* refers to *underage consumption of alcohol*. As the student constructs representations of additional claims, the tutor checks the diagram for inconsistent references and ambiguous variables,

which will initiate additional diagram tutoring. The strategy here is to have the student translate from her representation to the domain model representation by linking the diagram to text quoted from the report. Asking the student to translate between representations might be a waste of student time in some cases, but in this case, the student needs to link her diagram to the evidence in order to use the evidence during debate. This is one of the skills Policy World is explicitly trying to teach. In other words, the nature of the task allows the tutor to have the student translate between representations for free.

Synthesis

The fourth and final step of analysis is for the student to ***synthesize*** her overall beliefs about the causal relation between the two variables she is analyzing (Figure 5.10). To synthesize her belief, the student specifies whether she believes the first variable increases, decreases, or doesn't affect the second variable, and her confidence in that belief on a qualitative scale ranging from "completely uncertain" to "completely certain." While performing this synthesis step, the student can see her previous belief (if any) about the relation between these two variables, as well as the evidence that she's previously collected about these two variables.

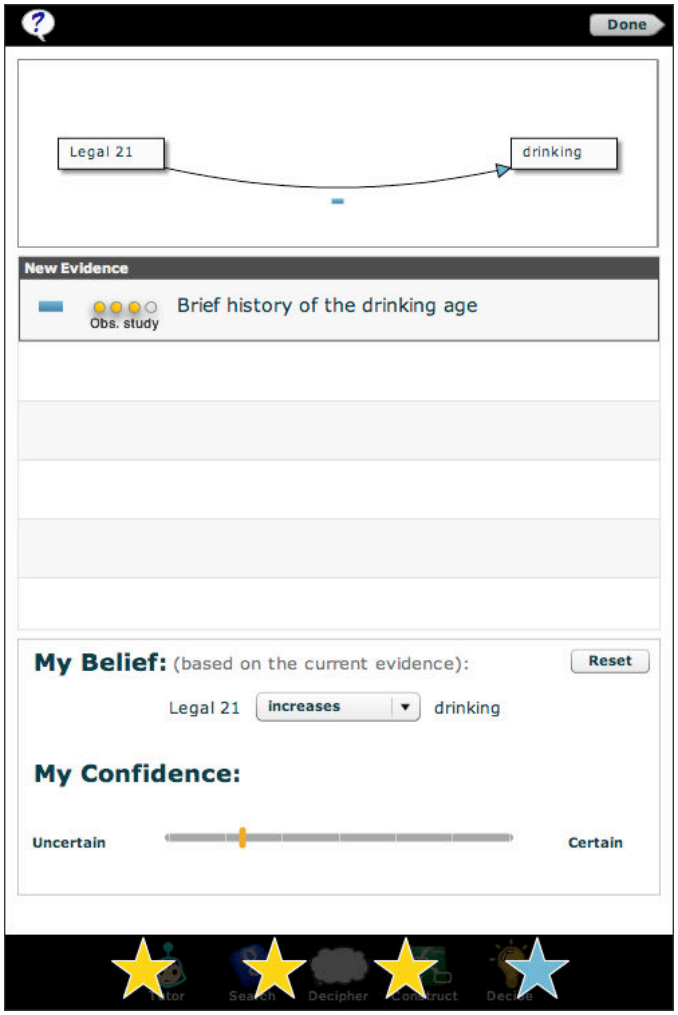


Figure 5.10. The interface for synthesizing beliefs in Policy World. The student specifies which causal relation between two variables (e.g., increase, decrease, negligible) is best supported by the evidence linked to all the arrows between those two variables. The student also specifies their confidence in that judgement.

Like evaluation, there is no precise quantitative specification of how to change one's beliefs after reading a new piece of evidence. However, Bayesian updating does provide us with constraints. There are two heuristics that should be followed: (a) the student's beliefs should move in the right direction, and (b) the student should not be dogmatic. Ideally, when synthesizing a new piece of evidence with preexisting evidence, we want students to move their beliefs in the correct direction, i.e., if the student encounters a new piece of empirical evidence that *legal age 21 doesn't affect drinking*, she should not then decrease her confidence in the belief that *legal age 21 doesn't affect drinking*, nor should she conclude that *legal age 21 decreases drinking*. Fortunately, Chapter 2 found that in general, students do move their beliefs in the correct direction.

Students should also not be dogmatic. That is, students' beliefs should mirror the balance of evidence. Unfortunately, Chapter 2 showed that students routinely violate the dogmatism constraint. While students do move their beliefs in the correct direction, they move it by such a small amount that their overall beliefs do not change. Thus new evidence is like drops of water on the hard rock of prior belief. Chapter 2 also found that students maintained very inaccurate pictures

of the evidence they had seen, sometimes even recalling that the majority of evidence supported their prior beliefs when the majority of evidence in fact contradicted their prior beliefs. Of course, if students could cite evidence for their prior belief, this might be rational, but in fact, unreported findings from that study indicated that students were either unable or unwilling to provide evidence for their prior beliefs.

To overcome the challenges created by ill-definition in synthesis, Policy World: (a) uses process constraints, and (b) teaches that students should have explicit evidence for belief, as in a court of law.

There are three process constraints (only the first two of which are currently used): (a) does the student move her new belief in the direction of the new evidence, (b) is the student's overall belief consistent with the overall evidence, and (c) is the student's overall belief consistent with her strength ratings of the evidence. For example, if the student receives a new piece of experimental evidence that *legal age 21 increases drinking*, then, if her prior belief is that *legal 21 decreases drinking* then she should decrease her confidence in that belief, or switch to a belief supported by the new evidence. If she already has the belief supported by the new evidence, then she should increase her confidence in that belief. For the second process constraint, Policy World calculates an evidence score for each causal relation as a function of the number of pieces of evidence for each claim weighted by a multiplier for each evidence type, where experiments have a higher multiplier than observational studies, which have a higher multiplier than cases, which have a higher multiplier than claims. If the student's belief about the causal relation does not match the causal relation with the highest evidence score, the tutor will warn the student that the evidence may not support that belief. For the third constraint, the tutor calculates a score for each causal relation by summing the student's strength ratings for each piece of evidence supporting the given causal relation. If the student's belief does not match the causal relation with the highest strength rating score, the tutor issues another kind of warning. Although the feedback provided by these three constraints might not be appropriate for a real policy analyst with a great deal of background knowledge, it does seem to provide appropriate advice in the context of the simplified problems presented in Policy World.

Policy World's second strategy for dealing with ill definition in synthesis is to tutor an "explicit evidence for belief" rule. In other words, the student isn't being asked to synthesize her belief about what she thinks is true, but her belief about what she thinks she can prove. The student may have any prior belief she wants, but she knows that she must explicitly cite evidence in order to win the debate, just like in a court of law. This is a *limit* strategy, in this case limiting what is taught by the expert model. Anecdotal comments from students indicate they understand the distinction and tension between the belief and proof. While we would like a version of Policy World that can also take students' background evidence into account, we would consider an initial version that succeeds at tutoring an explicit evidence rule a success.

Cross-examination

Depending on the tutoring mode, the student may enter a brief cross-examination phase after analysis where the tutor asks her to clarify the variable in her diagram. Cross-examination is unnecessary in cognitive mode, because the student's diagram will be fully and correctly linked to the evidence. Cross-examination is also unnecessary in game mode, because even if the student's diagram is not linked to the evidence, the tutor is not being expected to provide any tutoring specific to the student's representation. However, if the Policy World tutor *is* asked to provide tutoring

during the debate phase of the game, then it may be in the unfortunate position of having to interpret incorrect and uninterpretable student work.

If Policy World cannot make sense of the student's diagram (i.e., when the student performs poorly during stoic tutoring), the game resorts to the simple, if inglorious, approach of simply asking the student to explain her diagram. The tutor will tell the student what it thinks that her variables mean using the terms from the tutor's domain knowledge, and ask the student to confirm or correct the tutor's guesses. Note that this cross-examination step can also be incorporated into the debate phase of the game.

Cross-examination is simply another form of student translation that can be used in lieu of immediate feedback on linking during the diagram construction step.

Debate

In the first half of the problem/level, the student searches for and analyzes evidence. After search and analysis, the student moves to the final debate phase. In the debate phase, the student is asked to argue for a policy recommendation. The debate phase provides the motivation to engage in search and analysis and provides an opportunity to practice the *decision (via interpretation)* skills of the deliberation framework (Figure 2.2).

During the debate phase, the character playing the role of *judge* moderates a debate between the student and the character playing the role of the *opponent.* The four subtasks in the debate are: (a) to make a recommendation, (b) to explain the mechanism by which a recommendation affects the outcome, (c) to attack an opponent's mechanism, and (d) to provide evidence. Each debate has several rounds of attack and defense depending on the quality of the student's analysis, and how well she is able to use the products of her analysis. Table 5.1 shows some sample dialogue from a debate and Figure 5.11 describes the argument algorithm used to moderate the debate.

Recommendation

At the beginning of the debate, the character playing the role of the judge will ask the student for her policy recommendation, such as: *What should we do about the drinking age?* (Figure 5.12, upper left). Ideally, the student will then consult the diagram she's created to determine which intervention to recommend (Figure 5.12, upper right). She then chooses one intervention in the list of possible interventions to recommend (Figure 5.12, lower left). Finally, the character states the chosen recommendation (Figure 5.12, lower right).

Table 5.1

Sample Dialogue from the Debate Phase of Policy World

Character	Step	Dialogue
Recommendation subtask		
Mentor	1	What should we do about the drinking age?
Student	-	[Checks diagram created during the analysis phase]
	2	[Picks recommendation from list]
	3	I think we should repeal the 21 age drinking limit!
Mentor	4	[Decides that the recommendation is plausible, i.e., the recommendation is a variable that can be intervened upon, and moves to step 6]
Explain mechanism subtask		
Mentor	6	How will repealing the 21 age drinking limit affect drunk driving?
Student	-	[Checks diagram created during the analysis phase]
	7	[Selects mechanism using combo boxes]
	8	Repealing the 21 age drinking limit will decrease binge drinking which will decrease drunk driving! Also, repealing the 21 age drinking limit will decrease unsupervised drinking which will decrease binge drinking!
Mentor	9	[Decides that the mechanism is plausible, i.e., provides a causal chain from the recommendation to the outcome, and decides to move to step 11]
Evidence subtask		
Mentor	11	[Calculates which of the student's causal claims to attack and finds a weakness in the link from the 21 age drinking limit to binge drinking]
	12	What evidence do you have that the 21 age drinking limit increases binge drinking?
Student	-	[Checks diagram. Click on arrow in diagram corresponding to link between junk food and obesity to see the reports that the student has linked to the diagram.]
	13	[Picks report *Those who choose to drink are drinking more* from her list of evidence collected]
	14	This report shows that the 21 age drinking limit increases binge drinking!
Mentor	15	[Calculates quality of student's evidence and finds that the report cited does not support the student causal claim]
	16	That evidence is not convincing. In fact, that report does not contain any claims about the 21 age drinking limit and binge drinking at all.
		[Gives student a "strike"]

Note that in this training problem, the *judge* and *opponent* roles are both played by the mentor character.

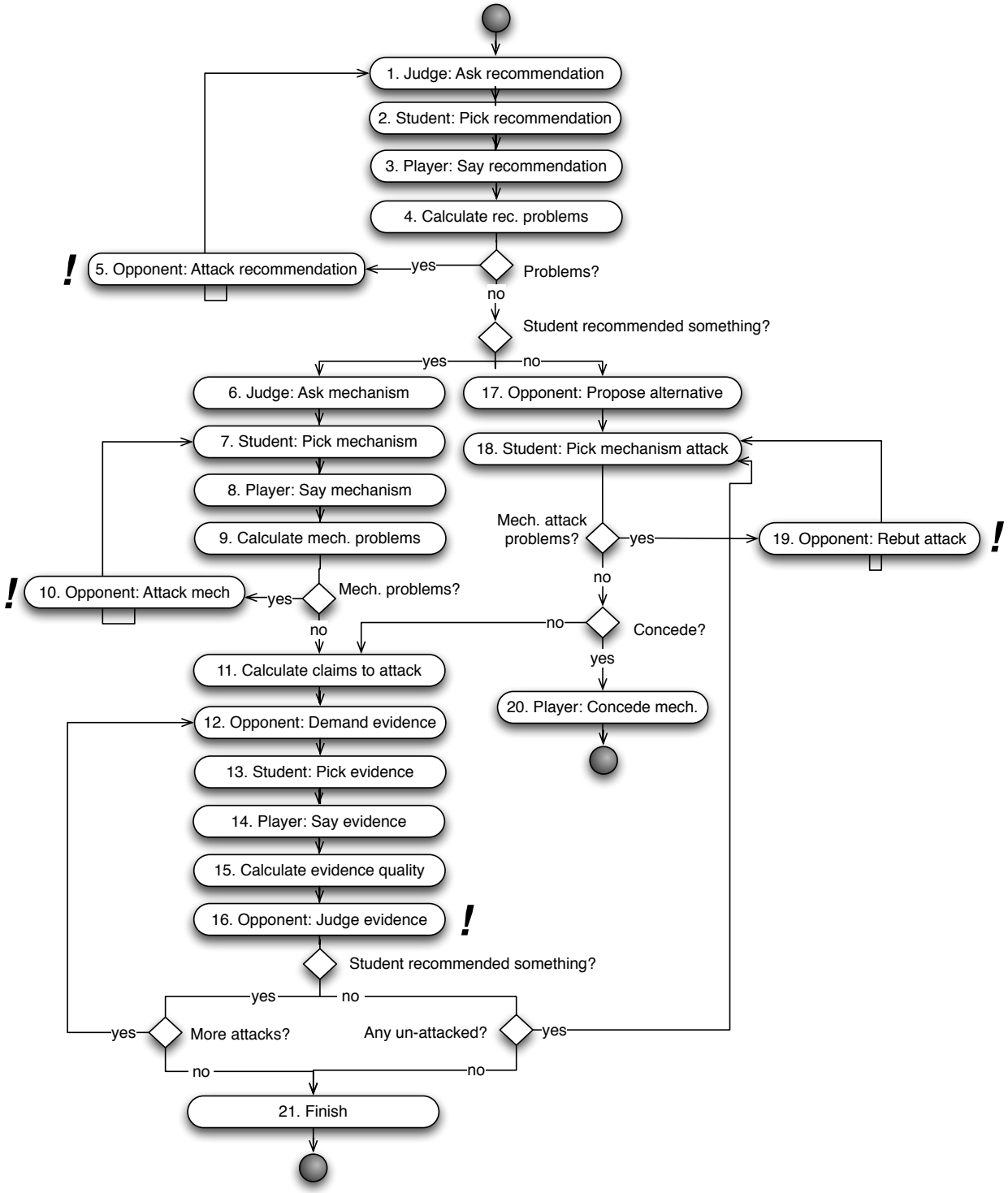


Figure 5.11. The argument algorithm used in the debate phase described as a UML activity diagram. Appendix D describes each of the 30+ argument moves that can be made during each activity. The student loses when he exceeds the maximum number of errors at states marked with an "!"

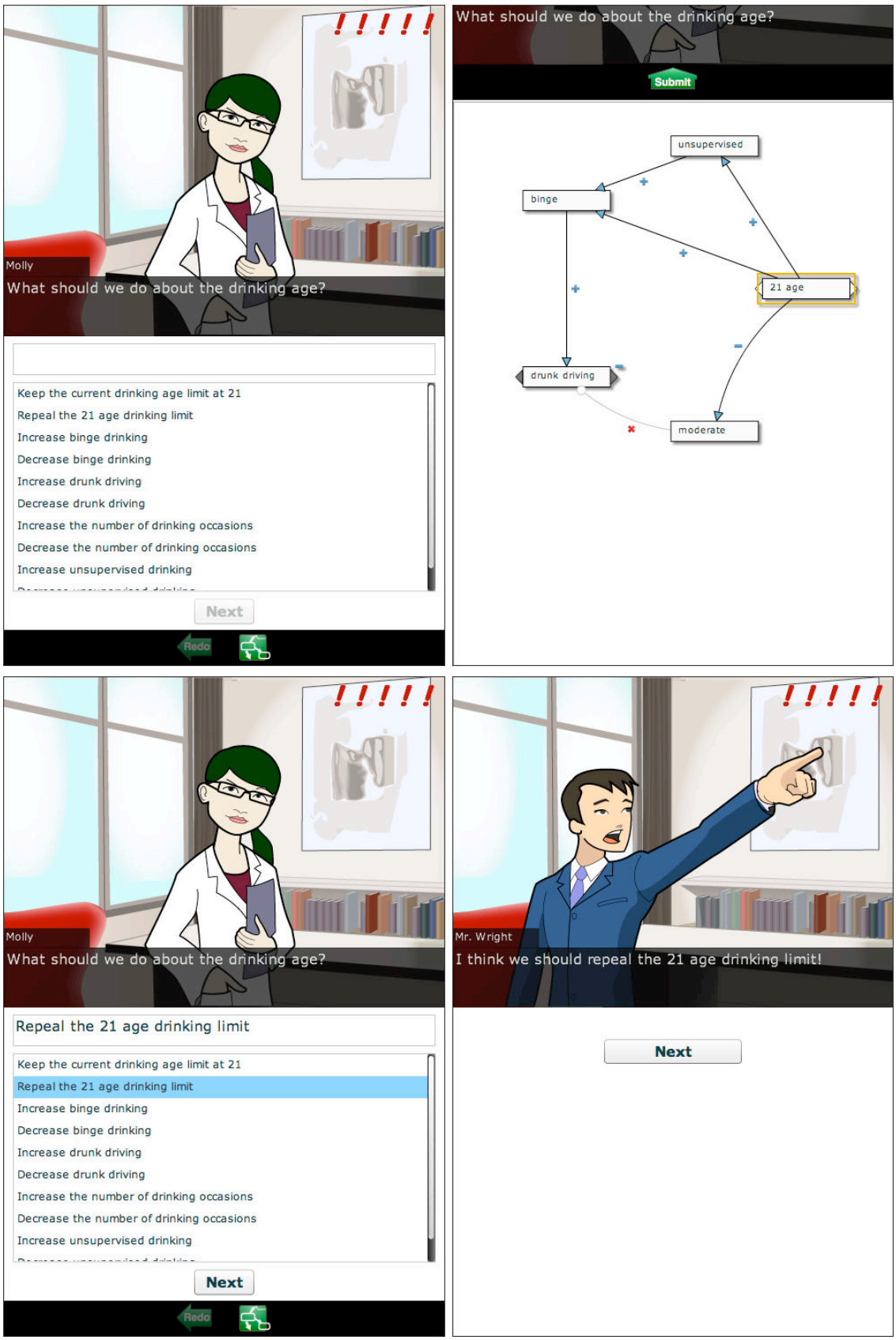


Figure 5.12. The interface for making a recommendation during the debate.

In terms of the deliberation framework, the pedagogical purpose of the recommendation subtask is to teach the *decision (via interpretation)* skills needed for using diagrams to make causal inferences (Figure 2.2). The rules for drawing inferences from a causal diagram are relatively well-defined assuming that the diagram does not contain contradictions or ambiguities. Although there are many sub-skills, the basic procedure for making a decision using the diagram is to: identify all the outcome variables that one seeks to change, identify which variables can be manipulated (i.e., the possible policy interventions), determine if there is a path (chain of arrows) from the intervention to the outcome, determine how to manipulate the intervention based on the signs of the arrows in the path, and explain the recommendation. This is essentially the "eliminate the cause(s) strategy" that experts use to solve policy problems (Voss, Tyler & Yengo, 1983), but supported here with causal diagrams based on the findings in Chapter 4.

Because the interpretation of causal diagrams is relatively well-defined, tutoring decision skills is relatively straightforward. However, even here issues of ill-definition affect tutoring. The first set of issues concerns whether the debate and whether decision tutoring should use the terms of the domain model or the terms from the student's representation. The second set of issues concerns how to provide tutoring when there are errors or ambiguities in the diagram.

First, consider whether the debate and tutoring should use the terms of the domain model or the terms of the student's representation. In Policy World, the debate is carried out using terms from the domain model, while tutoring takes place using terms from the student's representation. This requires the tutor to be able to map between the two representations. For example, the list of possible policy recommendations are described in terms from the domain model (Figure 5.12, left). The student may not have seen all of these terms, because they may be synonymous but not identical to the words of the text. For example the text of the student's report may contain a quote in an editorial like: *hamburgers make you fat* while the domain model encodes this variable as *obesity* which is displayed in the list of policy recommendations. During the debate phase, Policy World cannot necessarily conduct the debate using terms from the student's diagram, because the student may not have created a diagram in certain modes of tutoring such as the stoic and game modes. However, when the student is making a recommendation, if it is inconsistent with their diagram, then Policy World can provide tutoring using terms from the student's representation.

Figure 5.13 shows the some of the early steps of tutoring after the student has made a recommendation error. In Figure 5.13 the student has suggested that we *increase drunk driving*, which is not only incorrect, but not even plausible. The tutor responds by telling student to use her diagram (Figure 5.13, upper left). The tutor then asks the student to identify the policy outcome the student wants to affect (Figure 5.13, upper right). The student correctly identifies the outcome *drunk driving* which the tutor acknowledges (Figure 5.13, lower left). The tutor next asks the student: *What are the possible interventions?* (Figure 5.13, lower right). Tutoring of the "eliminate the cause(s)" strategy continues on in this manner using the representation created by the student. Table 5.2 shows the dialogue from the entire tutoring episode.

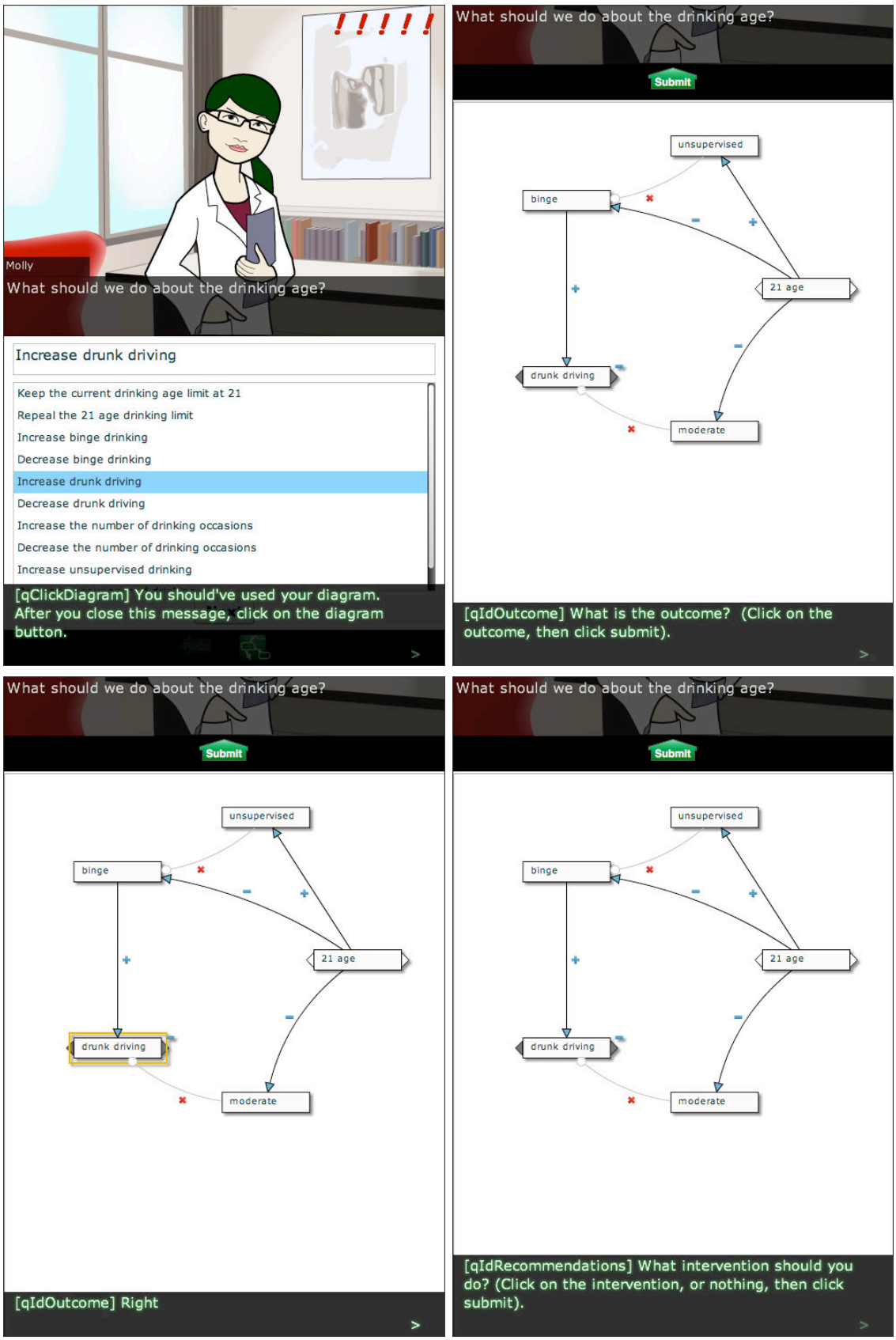


Figure 5.13. Screenshots of the initial steps of recommendation tutoring.

Table 5.2

Sample Recommendation Tutoring Episode Showing Character Dialogue and the Question Objects (Figure 5.17) Created by the Tutor

Character	Dialogue	Questions
Student	[Selects "increase drunk driving" a recommendation that is incorrect according to the student's diagram]	Recommendation constraint violated, pedagogical module adds *QRecommendation* to stack
Tutor	[interrupting the debate] That's not the right recommendation!	Tutor asks prompt from *QRecommendation*
Student	[Clicks to acknowledge message]	QReccomendation completes, adds QClickDiagram, QIdOutcome, QIdRecommendations, QIdInterventionManipulation and QDescribe to the stack
Tutor	You should've used your diagram. After you close this message, click on the diagram button.	Tutor asks prompt from *QClickDiagram*
Student	[Clicks to acknowledge message]	
	[Clicks diagram button]	QClickDiagram determines the answer correct and completes
Tutor	[Changes the screen to show the students diagram] What is the outcome?	Tutor asks prompt from *QIdOutcome*
Student	[selects *drunk driving* in diagram then clicks submit]	
Tutor	Right	*QIdOutcome* determines the answer correct, acknowledges student's answer and completes
Student	[Clicks to acknowledge message]	
Tutor	What intervention should you do? (Click on the intervention, or nothing, then click submit).	Tutor asks prompt from QIdRecommendations
Student	[Clicks to acknowledge message]	
	[Selects *21 age* in diagram, then clicks submit]	
Tutor	No	QIdRecommendations determines the answer incorrect, critiques the student's respomse and adds QUnallowableManipulation, QUndesiredOutcome, and QMaintainIntervention to the stack.
Student	[Clicks to acknowledge message]	
Tutor	Can you increase "21 age drinking limit"? [Yes/No]	Tutor asks prompt from QUnallowableManipulation
Student	[Clicks *No*]	
Tutor	Right	QUnallowableManipulation determines the answer correct, acknowledges student's answer and completes
Student	[Clicks to acknowledge message]	
Tutor	If you decrease "21 age drinking limit" will "drunk driving" decrease? [Yes/No]	Tutor asks prompt from QUndesiredOutcome
Student	[Clicks *No*]	
Tutor	Right	QUndesiredOutcome determines the answer correct, acknowledges student's answer and completes
Student	[Clicks to acknowledge message]	

Character	Dialogue	Questions
Tutor	How should you change "21 age drinking limit"? [Do nothing/decrease]	Tutor asks prompt from QMaintainIntervention
Student	[Clicks *Do nothing*]	
Tutor	Right	QMaintainIntervention determines the answer correct, acknowledges student's answer and completes
Student	[Clicks to acknowledge message]	
Tutor	What intervention should you do? (Click on the intervention, or nothing, then click submit).	Tutor asks prompt from QIdRecommendations
Student	[Clicks to acknowledge message]	
	[Selects nothing and clicks submit]	
Tutor	Right	QIdRecommendations determines the answer correct, acknowledges student's answer and completes
Student	[Clicks to acknowledge message]	
Tutor	[Switches to recommendations list] How would you describe your recommendation?	Tutor asks prompt from QDescribeRec
Student	[Selects *Keep the current drinking age limit at 21*, and clicks Next]	
Tutor	OK, try making your recommendation again	QRecommend determines the answer correct, acknowledges student's answer and completes
Student	[Clicks Next]	

For this recommendation tutoring to work, the tutor must be able to map between the student's recommendation in the debate, which is in the terms of the domain model, and the student's diagram, which uses terms created by the student. Recall that the link between the two knowledge representations has previously been created by the student in the diagram construction step. This resolves the first set of issues of ill-definition in tutoring the recommendation subtask.

The second set of issues of ill-definition in tutoring *decision (via interpretation)* concerns how to tutor incorrect or ambiguous diagrams. This issue can arise because the student has made an error (which is possible even in a well-defined situation), or because there are multiple possible representations due to the ill-defined nature of the policy problem (which is an issue in the domain model). In fact this is a minor issue, because all tutoring on diagram interpretation is relative to the student's diagram. In other words, if the student's diagram is inaccurate, then the tutor will teach the student how to accurately calculate the incorrect inferences implied by the diagram. This is like correctly solving an equation that has been set up incorrectly. With respect to multiple or ambiguous representations, the tutor simply makes the student commit to a particular representation during the synthesis phase. If the evidence about a certain causal relation is ambiguous, for example if there was one experiment showing that legal 21 law decreases drinking, and one experiment showing that legal 21 increases drinking, then the student could choose between either of the two causal relations without being criticized by the tutor. Once the student synthesizes the relation, she has committed herself to a particular diagrammatic representation. Diagram interpretation tutoring then proceeds from that representation.

As an aside, this tutoring of partially correct work could potentially lead to an interaction during the stoic mode of tutoring where the tutor's feedback and the debate feedback disagree. For instance, if the student's diagram incorrectly indicates that the *legal 21 age law* does not affect binge drinking, and that binge drinking increases drunk driving, then the student will recommend that she should not do anything to the drinking law. The tutor will look at the student's incorrect diagram and inform her that she has drawn a correct inference. However, the student's opponent will attack this recommendation as incorrect. The situation is akin to a game where the character walks into a dragon's cave armed with a toothbrush and receives tutoring on how to use the toothbrush – the character's actions might be correct with respect to toothbrush use, but the fatal error was during a previous decision. This situation could happen if stoic tutoring is configured to allow errors in diagram construction but not diagram interpretation, or perhaps in cases where there is ambiguous evidence that the intelligent debater might want to attack. This cannot arise in the cognitive mode of tutoring with unambiguous evidence, because the cognitive mode will not permit incorrect diagrams. This also cannot arise in the game mode, because the tutor will not intervene.

Explaining a mechanism

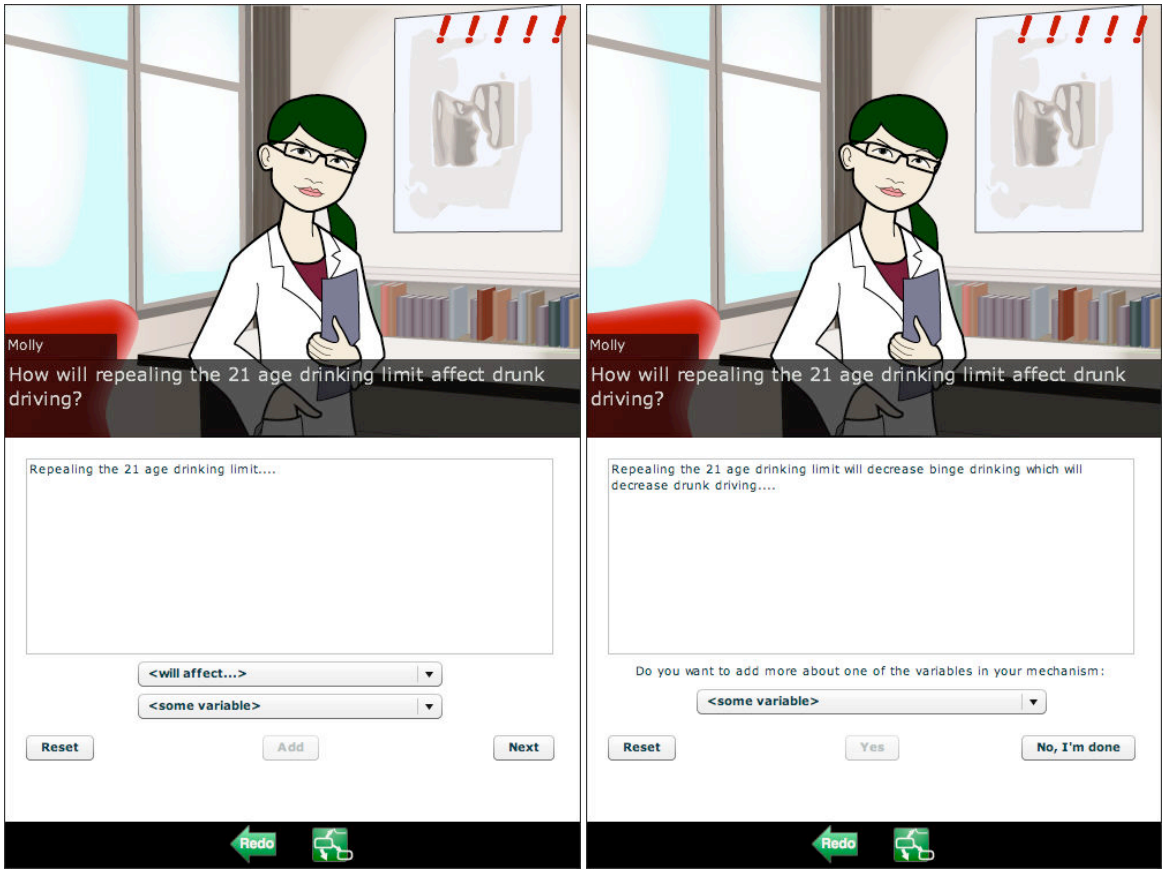


Figure 5.14. The interface for explaining a mechanism during the debate.

Once the student has made a plausible recommendation, the character in the judge role will ask the student to explain how her recommendation affects the outcome. For example, suppose the student has recommended that we *repeal the 21 age drinking limit* in order to decrease drunk driving. The student's reasoning might be that: (a) the the 21 age limit actually promotes binge drinking which

increases drunk driving, and (b) the 21 age limit increases unsupervised drinking which increases binge drinking. The student can construct a text-based representation of the mechanism using the interface shown in Figure 5.15. The text box in the middle of the screen in Figure 5.15 (left) shows the text of the student's initial recommendation: *Repealing the 21 age drinking limit…* The student can then chain causal effects onto this explanation. For example, the student can set the <*will affect...*> combo box to *will decrease*, and set the <*some variable*> combo box to *binge drinking*, which will create the text of the mechanism: *Repealing the 21 age drinking limit will decrease binge drinking...* The student can continue chaining causal effects with these combo boxes until there is a causal chain from the recommendation to the outcome like: *Repealing the 21 age drinking limit will decrease binge drinking which will decrease drunk driving.* When the student finishes describing a causal chain, she clicks the next button and is taken to the screen shown in Figure 5.15 (right). If the student wants to add an additional path, she can use the <*some variable*> combo in Figure 5.15 (right) to start the new path. For example, the student could set the combo to: *repeal the 21 age drinking limit*, starting a new path from which the student could then chain *decrease unsupervised drinking,* and *decrease binge drinking*.

When providing a mechanism, the student practices the same *decision (via interpretation)* skills practiced when making a recommendation. The only difference is that the output of the student's actions is a description of the whole mechanism rather than just the first variable in the mechanism (i.e., the recommendation). The challenges arising from ill-definition during the mechanism subtask are identical to those during the recommendation step, are addressed in the same way, and need not be described again.

Attacking an opponent's mechanism

The previous section on explaining a mechanism assumes that the student makes a recommendation like: *repeal the 21-age drinking limit* that increases or decreases some variable. However, in some cases there is no recommendation that will have the desired effect. In that case, the proper recommendation is to *do nothing*, for example, to maintain the current 21-age drinking limit. In this case, it makes no sense to ask the student how *doing nothing* will affect the outcome.

If the student recommends doing nothing, then the character playing the *opponent* role will suggest an alternate recommendation and mechanism, such as: *others claim that repealing the 21 age drinking limit will decrease binge drinking which will decrease drunk driving* (Figure 5.15, upper left). The student must then attack the alternate mechanism in three steps. First, the student is asked: *Do you think [this mechanism] is true?* (Figure 5.15, upper left). If the student answers, "No", she moves to the second step, if the student answers "Yes", then she loses the debate. In the second step, the student has to choose which of the causal claims in the alternate mechanism to attack (Figure 5.15, upper right). In the third step, the student has to propose the correct relation between the two variables in the causal claim attacked (Figure 5.15, lower left). Finally, the student's character will state the attack (Figure 5.15, lower right).

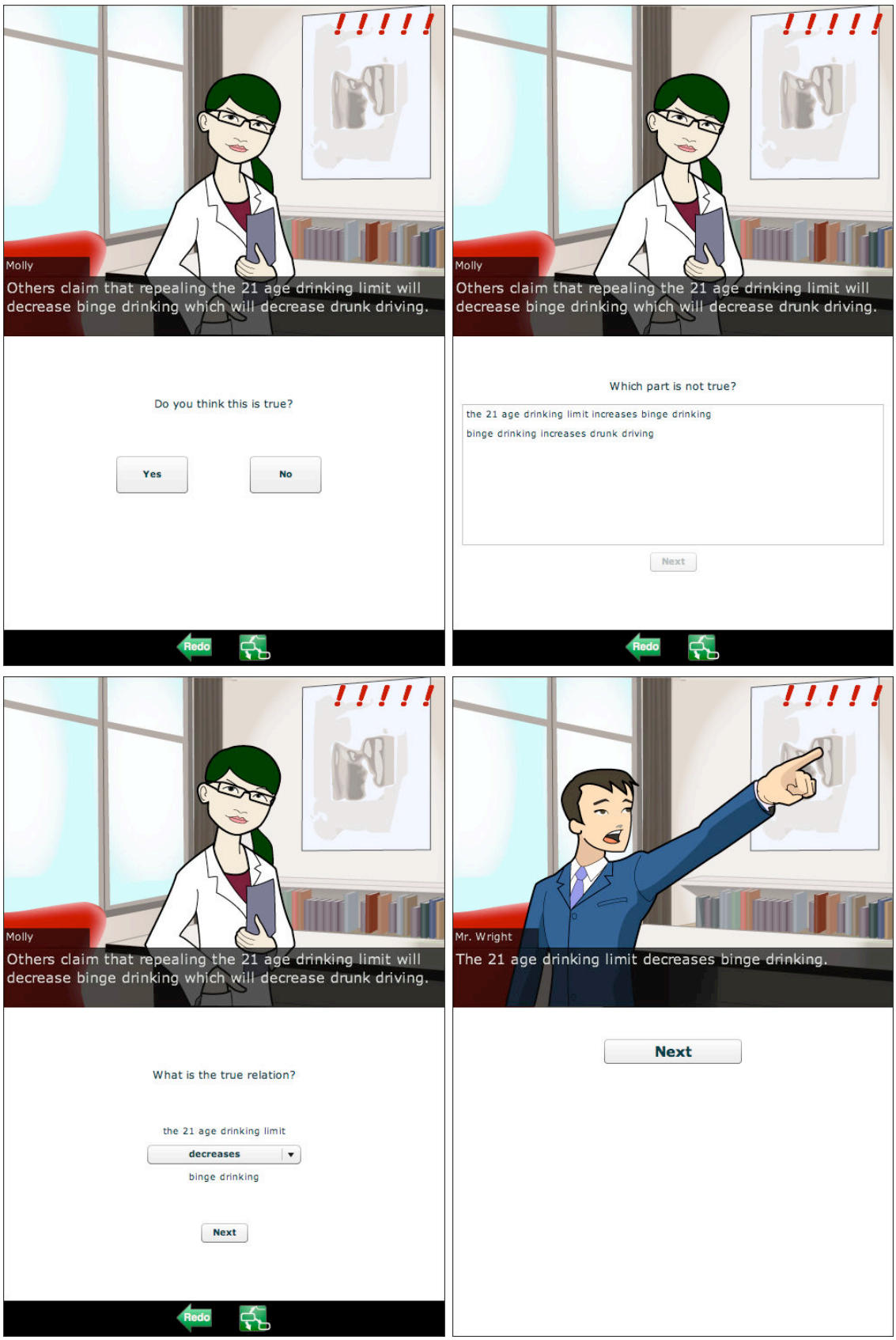


Figure 5.15. The interface for attacking an opponent's mechanism during the debate.

When attacking a mechanism, the student practices a slightly different set of *decision (via interpretation)* skills than those practiced in making a recommendation or explaining a mechanism, but the challenges of ill-definition and strategies for addressing them are the same, and need not be described again.

Evidence

Whether the student provides a mechanism, or attacks an opponent's mechanism, the character in the opponent role will select one of the claims in the student's mechanism, or the claim the student has attacked, and demand that the student provide evidence for her claims. For example, during the training level (where Molly also plays the role of the opponent), Molly may ask the student: *What evidence do you have that the 21 age drinking limit increases binge drinking*? (Figure 5.16, upper left). Ideally, the student will then check her diagram and click on the arrow corresponding to the claim that needs to be defended, in this case the arrow between *21 age* and *binge* (Figure 5.16, upper right). Clicking on that arrow will show the student all the pieces of evidence that she's linked to that arrow during the construction step of analysis (Figure 5.16, lower left). After closing her diagram, the student then cites those reports in defense of her claim (Figure 5.16, lower right). If the student cites stronger evidence for her claim than the opponent can cite against the claim, then the judge will accept the student's claim. When defending a mechanism, the student has to defend up to three causal claims before the judge will be convinced. Because Policy World attacks the student's mechanism with full awareness of the evidence the student has actually analyzed, the opponent may actually be tougher than a human opponent. When the student attacks the opponent's mechanism, the student only has to undercut one link in the opponent's causal chain.

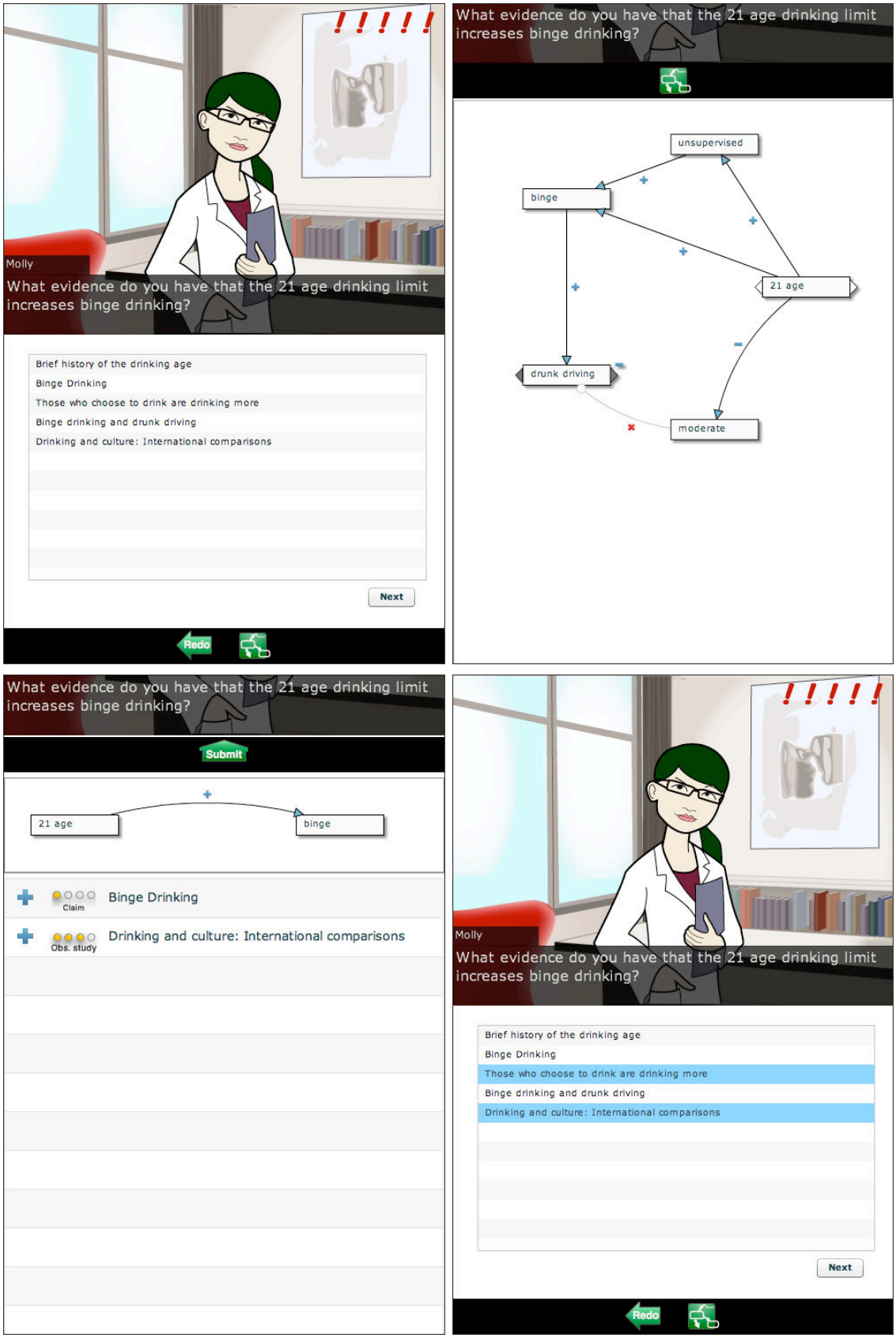


Figure 5.16. The interface for providing evidence.

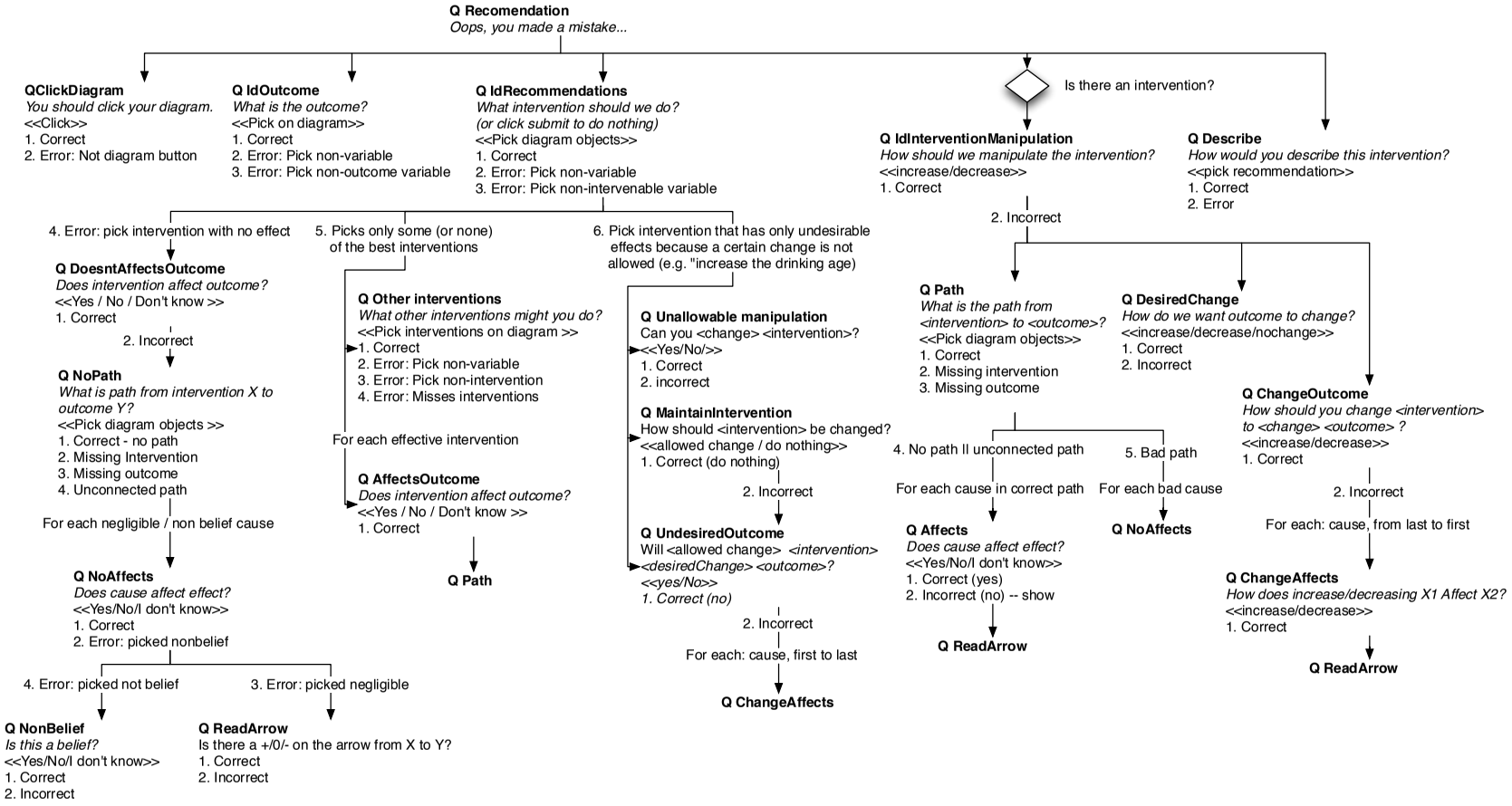
With respect to the deliberation framework described in Chapter 2, providing evidence requires just another subset of *decision (via interpretation)* skills. Providing evidence does create one additional challenge of ill-definition, specifically whether there is a correct answer to the problem. This is an issue for the tutor's domain model, but does not actually affect tutoring. The Policy World tutor is agnostic with respect to whether there is a correct answer in the domain model. In ambiguous cases, the tutor will not criticize the student so long as the student defends a policy position that is not weaker than another position.

However, for the sake of entertainment, each of the policy problems in Policy World has a winning position that allows the student to defeat the opponent. To accomplish this, Policy World *tilts* the evidence toward a specific policy recommendation. This means that one position will always have stronger evidence than the other positions, assuming the student can find it. The current version of Policy World polls students about their prior beliefs at the beginning of the game and stacks the evidence *against* the student's prior belief. So ironically, this tilting makes the problems more, rather than less, difficult. Furthermore, the evidence given to the student in Policy World is still much more ambiguous than in many science tutors in which all the evidence is consistent with the correct hypothesis.

The Socratic tutoring mode

The *decision (via interpretation)* skills that allow a student to make policy inferences from a causal diagram consist of a large number of non-verbal visual operations that are performed together and extremely quickly. Unlike in the *analysis* interface where the student performs an observable step for every skill traced by the expert model, the debate interface purposely does *not* require the student to explicitly perform each step. Forcing the student to perform each of these operations explicitly, including: identifying the possible interventions, identifying the outcomes, identifying the arrows that form a path between the outcomes, determining if the paths contradict, etc. would slow the debate phase to an intolerable crawl during the most critical part of the game. Instead of reifying each step during debate, Policy World uses a more dynamic, Socratic mode of tutoring.

Instead of making the student perform each step, the student is simply asked to provide the proper recommendation, mechanism, or evidence. If the student performs all the necessary diagram interpretation steps correctly, then Policy World moves on to the next subtask. If the student makes a mistake, then the tutor asks the first set of sub-questions. The tutor asks the prompt from the first sub-question and repeats the process recursively. A correct answer moves onto the next question; an incorrect answer spawns additional sub-questions. The Socratic tutor essentially performs a depth-first search, digging deeper until it locates the source of the student's error. If one thinks of the leaves of the question tree as the individual steps in a traditional cognitive tutoring interface, then one can see that the Socratic tutor provides a dynamic interface that skips past the steps the student can perform correctly and drills down to the steps where the student has problems. Figure 5.17 shows the tree of questions for recommendation tutoring, and Table 5.2 shows how they are used in a tutoring episode.



5.17. The set of questions used for tutoring decision, the process of making a recommendation based on the diagram. In the diagram above, each element represents 1 question. The name of the question is in bold, followed by a simplified version of the prompt given to the student. The third line (indicated by <<>>) describes the type of input the student is asked to provide. The numbered items describe all the possible ways the student's input can be evaluated. Arrows with numbered items leaving a question indicate that a subquestion is added to the question stack. This set of questions, if written in a traditional production system, would require slightly over 50 productions.

Conclusion

The first contribution of this chapter is to demonstrate how using a combination of several tutoring strategies: (a) to reify, (b) to limit, (c) to tilt, (d) to use weak model-tracing, (e) to use process constraints, and (f) to use student translation, can overcome problems of ill-definition. Table 5.3 summarizes the ways in which these strategies are applied in order to overcome the challenges created by ill-definition.

Table 5.3

Strategies for Addressing the Assistance Challenges Created by Ill-definition in Deliberation

	Challenge		Strategy	
Step	**Type**	**Description**	**Type**	**Description**
Question	*Expert model*	Policy problems can be reframed	*Limit*	Substantial reframing of the problem not possible
Search	*Input*	Monitoring search behavior	*Reify*	Explicit actions for for starting search, visiting site, finding report, analyzing report
	Domain model	Large search space of evidence	*Limit*	Search is not one of the primary learning goals, so this part of the task is restricted
	Expert model	No clear stopping rule for search	*Tilting*	The search space is minimized compared to real problems, so the tutor can advise searching for all sites / reports / claims.
Comprehend	*Input*	Monitoring comprehension	*Reify*	Explicit action for selecting causal claims
	Domain model	Messy information	*Limit*	Splitting multiple claims into two and removing duplicate claims
	Input	Reading comprehension	*Weak model*	Fuzzy sentence boundaries
Evaluate	*Input*	Monitoring evaluation of evidence	*Reify*	Explicit action for selecting among a limited set of evidence types and scale for specifying evidence strength
	Expert model	No rule for evidence strength	*Process constraint*	Enforce consistency and qualitative ordering
Construct	*Input*	Monitoring diagram construction	*Reify*	Explicit action for creating boxes. arrows and links
		Can't understand variable names	*Student-translation*	Student links diagram to evidence, mapping between student representation and domain model
Synthesize	*Input*	Monitoring causal beliefs	*Reify*	Explicit action for setting overall belief about the causal relation between two variables and quantifying confidence in that belief
	Expert model	No rule for judging body of evidence	*Process constraint*	Enforce moving belief in right direction and explicit evidence

	Challenge		Strategy	
Step	**Type**	**Description**	**Type**	**Description**
	Input	Students background beliefs	*Limit*	Debate favors explicit evidence over background knowledge
Decide	*Input*	Monitoring diagram interpretation	*Reify*	Explicit action for choosing among list of possible interventions and selecting diagram elements during diagram interpretation tutoring
	Feedback	Can't describe students input in natural language	*Student-translation*	Debate uses terms from tutor's domain model and tutoring uses student term, relying on linking in diagram construction step to map between the two
	Domain model	Want to have a winning answer	*Tilting*	Stack the evidence to favor one policy position over the others

Overcoming these challenges of ill-definition paves the way for the second contribution: a system that can tutor deliberative argument. Policy World restricts the range of argument to causal arguments and uses the causal diagrams constructed and synthesized by the student to make student's position explicit and machine-readable. By restricting and defining the problem in combination with the argumentation algorithm, Policy World is able to debate the student in a realistic (albeit limited) manner.

Policy World was also designed to combine the best qualities of tutoring systems with the best qualities of games. Such a system requires a more flexible pedagogical approach that can provide assistance spanning the continuum between the more explicit, immediate feedback of the tutor and the more situational, less frequent assistance of the game. The third contribution of this chapter is an architecture for a pedagogical module that can dynamically switch between direct, cognitive, Socratic, stoic, and game-based modes of feedback. This pedagogical module makes it feasible to combine an intelligent tutor with a game-based inquiry environment and provides a platform for experimentation used in Chapter 6.

Policy World shows how the combination of a cognitive framework for deliberation, a diagram-based inquiry environment, an argument algorithm, and a Socratic tutor can provide argumentation tutoring in the domain of public policy. This system should be of interest not only to those directly concerned with teaching public policy and civics, but to other disciplines that require evidence-based arguments about causal systems such as science, history, business, as well as those disciplines in which students represent arguments in diagrammatic form such as law, philosophy, HCI, etc. This chapter demonstrates that it is *possible* to provide argumentation tutoring. We must now test whether this approach to argumentation tutoring is *effective*.

6. Combining games with intelligent tutors to improve learning and motivation

Summary: Combining educational games and intelligent tutors leads to a conflict in how best to provide assistance: games offer minimal assistance and impose penalties, whereas tutors provide more assistance and allow students to correct errors. The lack of empirical work comparing games and tutors provides little guidance in how best to resolve the conflict. In this study, I investigated whether game-based or tutoring-based assistance was more effective at increasing learning and interest in a policy reasoning task. In this laboratory experiment, 78 university students played one of two different on-line versions of the Policy World game. The *game* version of Policy World provided primarily minimal feedback with penalties. The *cognitive game* version of Policy World added a tutor that provided knowledge-based feedback on each step and required immediate error-correction. The experiment used a randomized, controlled, 2-group, between-subjects design. Log data from Policy World was used to construct learning measures for each step of policy reasoning, e.g., comprehension, evaluation, diagram construction, synthesis, and decision. The Intrinsic Motivation Inventory was used to assess interest. The results showed that *cognitive game* version of Policy World increased learning of analysis and motivation more than the *game* version. These findings suggest that we can increase the effectiveness of educational games with a tutoring-based approach to assistance.

In the previous chapter, I described how Policy World overcomes the technical obstacles involved in analyzing policy reasoning in order to provide assistance in an ill-defined domain. Policy World's inquiry environment (Chapter 5) was designed based on the cognitive model of deliberation (Chapter 2). It focuses on the learning challenges that are caused by bias, especially those that arise primarily during analysis as opposed to search (Chapter 3). And it uses causal diagrams as a central learning tactic, because they improve learning and performance (Chapter 4).

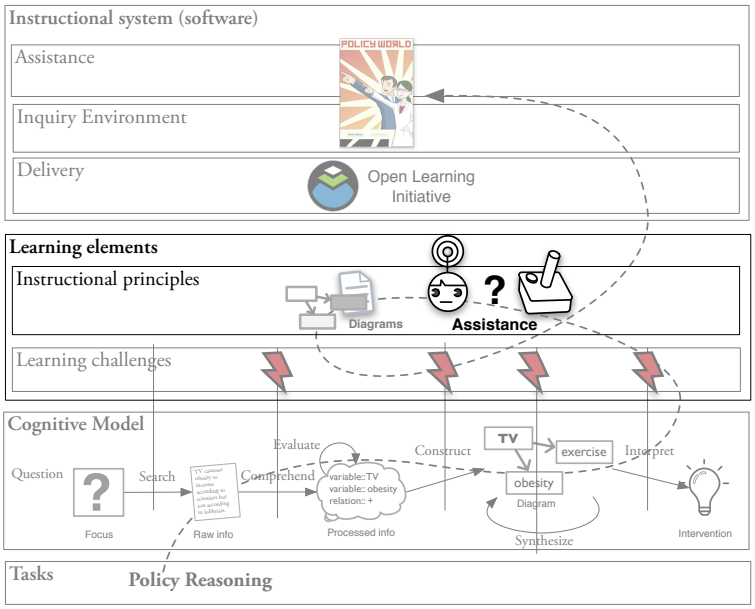


Figure 6.1. Chapter 6 considers a question of instructional principles: how best to provide assistance when combing tutors and games.

While the design of Policy World was strongly influenced by the cognitive work described in the previous chapters, the findings at the learning elements level do not sufficiently determine the design of an instructional system. One of the most significant, untested design decisions was to combine an educational game with a tutoring system. This chapter considers the fifth research question: *How can we best provide assistance when combining tutors and games?*, a problem at the level of instructional principles (Figure 6.1).

Unfortunately, our approaches to designing educational games and intelligent tutors lead us to conflicting sets of instructional systems. Some conflicts such as whether or not to use fantasy are easy to resolve. Whereas video games typically use fantasy environments and narratives, tutors usually do not. Anecdotally, the very minimal fantasy context used in the search and analysis experiment described in Chapter 3 seemed to increase students' engagement relative to the no-fantasy context of the diagram studies in Chapter 4. We also know that adding fantasy contexts to educational games can improve both learning and interest (Cordova & Lepper, 1996). So embedding a tutor in a game-like fantasy context seems like a straightforward and likely effective design decision.

Other differences between tutors and games, such as how to provide assistance, are more difficult to resolve, because we have to choose one approach or the other and our choice is likely to affect learning and interest. Games and tutors differ in how they provide assistance. Tutors such as Cognitive Algebra, Andes, and Steve, typically provide step-level, knowledge-based teaching feedback that explicitly explains errors and principles either directly or via hint messages (VanLehn 2006). Tutors also often allow students to immediately correct their errors. This is in stark contrast to games. Games rarely provide knowledge-based feedback and, instead of allowing immediate error correction, games typically impose penalties for making mistakes such as decreasing *health* or *death*. Although tutors using mastery learning may impose indirect penalties by requiring students to complete more problems, they do not impose an immediate, direct penalties such as forcing the student to restart a problem the way that players who die in a game must restart the level. These different design approaches to assistance lead to very different experiences. What is not clear is whether games' approach to assistance is a necessary part of their allure, or if we can create systems that look like games but assist like tutors and are just as interesting.

For the purposes of contrasting these approaches to assistance, I will define *game-based assistance* as a combination of minimal and situational feedback that is not necessarily provided at the step level but gives immediate penalties for errors. I will define *cognitive game-based assistance* as the baseline game-based assistance augmented with additional cognitive tutoring. Specifically this includes the addition of: (a) knowledge-based feedback provided on every step, and (b) the requirement to immediately correct errors. By situational feedback I mean responses to student actions that cannot be removed from the problem solving environment without changing the task. For example, if a student debater provides evidence that supports his opponent's claim rather than his own, the opponent will argue that evidence supports the opponent's claim. The opponent's response is intrinsic to the debate task and provides this student with indirect feedback that the student has made an error. Situational feedback is indirect, because it is up to the student to infer what the situational response says about the correctness of his action. For example, the opponent might claim that the student debater's evidence supports the opponent's claim, but the opponent could be incorrect or lying, (whereas the student can assume that an intelligent tutor's explicit, knowledge-based feedback is correct). By

knowledge-based feedback I mean error-specific feedback, teaching feedback, and hints. Here knowledge-based feedback refers to any feedback that includes explicit instruction beyond minimal, correct/incorrect, error-flagging

One possible objection to these definitions is that the constructs of game and tutor cannot be defined. The categories of game and tutor are imprecise and overlapping, so it is possible to imagine games and tutors that that do not strictly adhere to this distinction we've made between *game* and *cognitive game.* However, it is certainly the case that the vast majority of video games offer less than step-level, knowledge-based feedback and that they include penalties; think of prototypical examples such as Halo, Legend of Zelda, and Guitar Hero. Likewise, the vast majority of tutors offer step-level feedback which is very often knowledge-based, and the vast majority allow students to correct errors; think of prototypical examples like Cognitive Tutor Algebra, Andes, and Steve. If these definitions do not describe all games and tutors, they certainly describe a great many.

A second objection to this definition is that the constructs are confounded, because they vary multiple attributes, for example the type of feedback and the use of penalties. However, if one accepts the definition, then at the level of the construct, only 1 attribute is varied: tutor (cognitive game) or no tutor (game). By analogy, a comparison of cars to airplanes varies by only one attribute at the level of the construct, even though the constructs of car and airplanes vary by multiple sub-attributes (numbers of wheels, number of windows, etc.). This does not make a comparison of cars to airplanes confounded. One simply has to be careful to draw inferences about the comparison at the proper level. In any case, the current swarm of activity on designing intelligent tutors and educational games warrants research into controlled comparisons between the two, however imperfect.

Intuitively, we might expect tutors to be more effective at increasing learning, because the principles upon which they are based have been derived from decades of empirical work (Koedinger & Corbett, 2006), and because of the empirically demonstrated benefits of immediate, knowledge-based feedback with immediate error-correction (Corbett & Anderson, 2001). On the other hand, situational feedback, and delayed intelligent novice-feedback, similar to that offered by games, can be just as effective or even more effective at promoting learning as immediate, knowledge-based feedback (Nathan, 1998; Mathan 2005). Intuitively, we might expect the game to be more fun, because it gives the player more autonomy and the satisfaction of winning. On the other hand, excessive floundering is not fun, and the additional assistance offered by the tutor might be welcomed by a struggling student. These competing intuitions and potential tradeoffs form the core of the assistance dilemma (Koedinger & Aleven, 2007).

While there is growing interest in educational games, the empirical research gives us little guidance in predicting how the game might fare against a tutor. An extensive review of the empirical literature on educational games by Hays (2005) concluded that "there is no evidence to indicate that games are the preferred instructional method in all situations," i.e., that there is a dearth of randomized-controlled experimental studies comparing games with other kinds of instruction such as cognitive tutors.

Research question

The purpose of this study was to test the effectiveness of combining tutors with games as an instructional approach and the effectiveness of Policy World for teaching policy argument. The study compared the effects of different types of *assistance* on *learning* and *interest.* The assistance variable compared a *game* and a *cognitive game* version of Policy World. In the *game* version (no tutor), the student received only a baseline level of assistance that included: situational feedback such as the game characters' dialogue, minimal-error flagging via the scoreboard, and penalties for making errors, such as having to restart the level. In the *cognitive game* version (with tutor), the student received additional knowledge-based feedback on every step and was required to immediately correct errors. The *learning* variables consisted primarily of students' learning on each stage of the cognitive model of deliberation including: search, comprehension, evaluation, diagram construction, synthesis, and decision. The *interest* variables were measured using the Intrinsic Motivation Inventory, a validated instrument with sub-scales for: interest/enjoyment, perceived competence, effort/importance, pressure/tension, perceived choice, and value/usefulness.

These variables allow us to pose several competing hypotheses about the benefits of combining tutors and games:

1. **Game hypothesis**: game-based assistance will increase learning and interest.
2. **Cognitive game hypothesis**: tutor-based assistance will increase learning and interest.
3. **Assistance tradeoff**: game-based assistance will increase interest, while tutor-based assistance will increase learning.

We might predict results consistent with the assistance tradeoff hypothesis given the conventional wisdom on games and the research on cognitive tutors. The pedantic feedback of the tutor could decrease perceptions of choice and thus interest. The lower assistance of the game might not have as strong an impact on learning. However, the cognitive game might increase students' interest more than the game if it increases their feelings of competence. The game might increase students' learning the most if it provides effective situational or intelligent-novice feedback. Furthermore, it's hard to predict how differences in interest will impact learning.

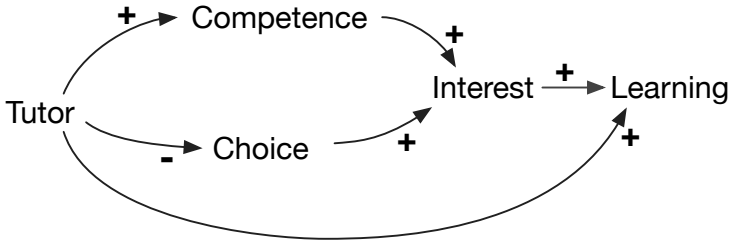


Figure 6.2. Possible mechanisms by which assistance may affect learning and interest.

Whether the results support the game, cognitive game, or assistance tradeoff hypothesis depends on the relative strength of competing causal paths. We may find support for the game hypothesis if the less explicit assistance of the game does not harm learning or competence and students perceive more choice, thus increasing interest, thus increasing learning. We may find support for the cognitive game hypothesis if the tutor improves learning, and the increased feeling of competence outweighs any perceived decrease in choice. We may find support for the assistance tradeoff hypothesis if the

tutor increases learning, but the perceived decrease in choice outweighs any increased perception of competence, thus harming interest.

Method

Population / Setting

Seventy-eight university students were recruited through an on-line participant database and flyers distributed around campus. The study was conducted on-line. After the study, students were compensated $20 for completing the study, plus an additional $5 for passing posttest 1, plus an additional $5 for passing posttest 2.

Design

The study used a two-group, randomized, controlled, experimental design that compared a *game* (*no tutor*) version of Policy World to a *cognitive game (with tutor*) version of Policy World. Both versions of Policy World offered a baseline level of feedback appropriate to the game-like inquiry environment. This included: minimal error flagging during analysis via a scoreboard, and situational feedback in the form of the characters' dialogue during debate. The two versions differed on whether or not they included cognitive tutoring. In the *game* (no tutor) version, students received only baseline assistance. When they made a mistake they received a penalty that required them to redo work. In the *cognitive game* (with tutor) version, students received additional step-level knowledge-based feedback on each step. Furthermore, instead of receiving a penalty, students were required to immediately correct errors. In other words, the tutor always provided hints while the game let the student die.

To make the distinction concrete, consider two examples of the game-based and tutor-based assistance provided by the different versions of Policy World during analysis and debate.

Assistance during analysis

In the analysis phase of problem solving students were asked to analyze evidence in the form of short, 3-5 paragraph, newspaper-like reports based on real articles from sources like the *New York Times* and PBS's *Frontline.* After searching for these reports using a fake Google interface, the student's *reports screen* displayed the list of reports to analyze. To begin analyzing a report, the student clicked on one of the reports in the list. The student then had to perform a series of four steps. First, the student *comprehended* the report by selecting a causal claim from the text. Second, the student used combo boxes to identify the evidence type and strength of the causal claim. The possible evidence types were: *experiment, observational study*, *case*, or *claim*, while evidence strength was rated on a 10 point scale with the labels: *none*, *weakest*, *weak*, *decent*, *strong*, *strongest.* Third, the student *constructed* a diagrammatic representation of the causal claim using boxes to represent variables, and arrows to represent an *increasing, decreasing,* or *negligible* causal relationship between the two variables. Fourth, the student *synthesized* his overall belief about the causal relationship between the two variables based on all the evidence found about those variables up to that point. The synthesis step required the student to specify which causal relationship between the two variables was best supported by the evidence, and his confidence in that relationship on a 100 point slider from *uncertain* to *certain.*

In both the *game* and *cognitive game* versions of Policy World, the analysis scoreboard would flash a large red star if the student made a mistake on any of the four steps of analysis. If the student made a mistake in the *game* version of Policy World, he had to restart the analysis of that causal claim. For example if the student made an error while constructing a diagrammatic representation, he received a red star and was sent back to the reports screen, losing his work on the comprehension and evaluation step for that claim. If the student made an error in the *cognitive game* version, he received knowledge-based feedback about the error and was required to correct the error. For example, if the student made an error while constructing a diagrammatic representation, he received a red star and an explicit feedback message such as:

> This quote is about the effect of "21 age drinking limit" on "moderate drinking." In your diagram, the box "drinking" represents "moderate drinking." However you created another box "forbidden fruit" to represent "moderate drinking." You should remove the box "forbidden fruit," and make your arrow point to the box "drinking."

The student then had to reattempt that step.

To characterize the similarities and differences between the *game* and *cognitive game* during analysis: (a) both versions provided minimal (error-flagging) feedback via the scoreboard stars, (b) both versions provided feedback on each step, (c) only the *cognitive game* provided knowledge-based hints, and (d) the *game* penalized students whereas the *cognitive game* required students to immediately repair errors.

Assistance during debate

During the debate, the student performs four types of subtasks. The first subtask is to provide a policy recommendation, such as: *We should repeal the 21 age drinking limit.* This task ideally requires the student to consult the diagram he created during the analysis phase before selecting a policy recommendation from a list of possible recommendations. The second subtask is to provide a mechanism explaining how the recommendation affects some desired policy outcome, such as: *Repealing the drinking limit will decrease binge drinking which will decrease drunk driving*. The student ideally consults his diagram before constructing an explanation using different sets of combo boxes. The third subtask is to attack an opponent's mechanism and is only required when the student recommends doing nothing. In that case, the student's opponent will make a policy recommendation; the student will select one causal claim in the opponent's mechanism to attack, and provide an alternate causal relationship between the two variables in the opponent's causal claim. The fourth subtask is to provide evidence for a causal claim by citing reports that support that claim from the list of reports collected by the student. Ideally, the student will consult his diagram by checking the list of reports the student has connected to each causal arrow during analysis. If the student has provided an explanation of a mechanism, the opponent will attack up to three causal claims in that mechanism before the student wins the debate. If the student attacks an opponent's mechanism, the student only has to provide evidence for one attack (which invalidates the opponent's entire causal chain). In the debate phase, the student is allowed to make 5 mistakes before losing the debate.

For the first three subtasks, making a recommendation, explaining a mechanism, or attacking an opponent, both the game and the cognitive game provide situational feedback via the character dialogue if the student makes a completely implausible move. For example, if the student suggests

that we decrease parental permissiveness (which is not possible) in order to decrease obesity, the judge might object: "That's not a valid intervention! We don't have control over how parents raise their children." When providing a recommendation, some of the variables can be intervened upon while others cannot. The *legal 21 age drinking age* can be repealed, or left in place. Both recommendations are plausible, even though only one or the other can be successfully defended given the available evidence. Other variables cannot be intervened upon, for example decreasing the outcome *drunk driving* is not a plausible recommendation. If the student makes an implausible recommendation, one of the characters will object, and the student will receive a "strike"(where 5 strikes results in losing the game). The characters will not object if the student makes an incorrect, but plausible recommendation. Similarly, there are also implausible mechanisms, such as the mechanism where the recommendation is not connected to the outcome. There are also implausible attacks, such as when the student attacks a causal claim in the opponent's mechanism by providing a causal claim identical to that provided by the opponent. When providing evidence to defend a causal claim, the student is supposed to cite reports that contain the causal claim being defended. If the student cites stronger evidence than the opponent's counter evidence, the characters' dialogue indicates that the student's evidence wins.

This type of feedback provided by the debate characters is *situational,* i.e., it is an inherent part of the debate environment. This feedback is also *not* at the step level in the sense that the student might have to perform multiple diagram interpretation steps before producing a recommendation. Furthermore on the first three subtasks, the situational feedback indicates only that the student's action is plausible, not necessarily that the action is correct. Both the *game* and *cognitive game* provide this baseline level of feedback.

If the student makes an error in the *cognitive game* version, the tutor immediately intervenes with a series of Socratic questions that teach the student how to use his diagram to perform the relevant debate subtask. Furthermore, instead of receiving a strike, the student is required to immediately repair his error. The type of feedback provided by the tutor is knowledge-based, step-level feedback that requires immediate error-correction.

To characterize the similarities and differences in assistance during debate: (a) both the *game* and *cognitive game* provide situational, error-specific feedback when the debate characters object to the student's move, such as when the student presents weak evidence, (b) the *game* version does not necessarily provide feedback on every step, whereas the *cognitive game* version does, (c) the *cognitive game* provides knowledge-based feedback whereas the *game* does not, (d) the *game* allows students to fail the debate and restart the level, whereas the *cognitive game* requires students to immediately repair errors.

Table 6.1 summarizes the differences in the types of assistance provided by the game (which provides baseline feedback) and the tutor (which provides both the baseline and additional tutoring feedback).

Table 6.1

Sample Baseline and Tutoring Feedback on Each Stage of Deliberation

		Sample Assistance	
Stage	**Possible error**	**Baseline**	**Tutor**
Search	Search too few reports	-	“You didn't find enough reports.”
Comprehend	Select non-causal claim	Red star	“That's not a causal claim”
Evaluate	Identify incorrect evidence type	Red star	"That's not the right evidence type"
Construct	Constructs arrow with incorrect relation	Red star	"In this quote, the first variable increases the second, but in your arrow the first variable decreases the second."
Synthesis	Shifts confidence in incorrect relation	Red star	"No, the new evidence shows that the first variable 'junk food' increases the second variable 'obesity'. But you increased your confidence in a different relation."
Recommendation	Selects non-intervention	Judge exclaims that this is not a valid intervention and gives student a strike.	Socratic questions on causal paths between recommendation and outcome.
	Selects possible (but not best) intervention	-	Socratic questions on causal paths between recommendation and outcome.
Mechanism	Select incoherent mechanism	Judge makes exclamation about problem with mechanism and gives student a strike.	Socratic questions on causal paths between recommendation and outcome.
	Selects plausible (but not best) mechanism	-	Socratic questions on causal paths between recommendation and outcome.
Evidence	Cites irrelevant evidence	Judge exclaims that this evidence is irrelevant and gives student a strike.	Socratic questions on using diagram to identify evidence for student’s claim.
	Cites winning, but not all of the relevant evidence	Judge says that the evidence is convincing.	Socratic questions on using diagram to identify evidence for student’s claim.
Mechanism attack	Attacks a cause that is not in opponent's mechanism	Judge exclaims that cause was not mentioned by opponent and gives student a strike.	Socratic questions on using diagram to identify conflicting claims.
	Attacks a cause	-	Socratic questions on using diagram to identify conflicting claims.

Note. Policy World recognizes close to 100 types of errors – see Appendix D and Appendix E for a detailed description of errors and feedback.

Procedure

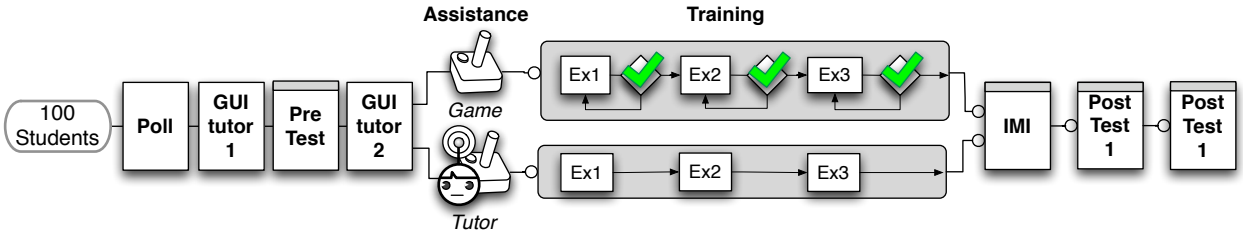


Figure 6.3. Experimental procedure using a two-group, between-subjects design.

At the beginning of the experiment, students took a **poll** on their policy beliefs about each of the 6 policy problems in the study. Each question asked the student about his position on a particular topic, e.g., "Do you think the legal-21 drinking age should be lowered to 18 years?" [yes/no]. The program then generated the evidence provided in each problem so that it contradicts the student's initial position, e.g., if the student thought we should repeal the legal-21 drinking age, then the majority of evidence supported maintaining legal age at 21.

Next, students were given a brief tutorial (**GUI Tutor 1** in Figure 6.3) that explained how to use the debate interface needed to complete the pretest. The tutorial consisted of direct instruction telling the student how to use the interface to provide answers given by the tutor, e.g., "Use the list below to recommend that we decrease smoking."

Students were then given the **pretest**. The pretest, posttest 1 and posttest 2 problems were counterbalanced. The topics of the problems were: (a) junk food advertising and childhood obesity (13-15 causal statements), (b) health care (8-9 causal statements), and (c) cap and trade (9-10 causal statements). During the pretest, the analysis tools for comprehension, evaluation, construction and synthesis were not available. Students were allowed to search for as many or as few reports as they liked before proceeding to the debate. The only feedback students received was the situational feedback intrinsic to the debate.

After completing the pretest, students received a second tutorial (**GUI Tutor 2** in Figure 6.3) that explained how to use the analysis tools. This tutorial also consisted of direct instruction telling the students how to use the interface, e.g., "Add a box to the diagram then double click it and type 'smoking'."

Students were then randomly assigned to either the ***game*** *(no tutor)* or the ***cognitive game*** *(with tutor)* condition. Each group completed 3 **training** problems on: (a) video game violence, which had 3 causal statements, (b) the legal-21 drinking age, which had 12 causal statements, and (c) the methamphetamine epidemic, which had 8 causal statements. As described in the previous section, *game* students received only the baseline assistance during the analysis phase, were allowed to proceed to the debate without analyzing all claims, and received only situational feedback during the debate. *Cognitive game* students received baseline assistance, received knowledge-based explanations during analysis, were required to analyze all claims before proceeding to the debate, and were given Socratic tutoring during the debate. In the *game* condition, it was possible for students to lose the debate, in which case they had to retry the problem at least once. After retrying the problem for the second time, the student had the option to retry the problem or proceed to the next training problem. The

tutor required immediate error correction, so *tutor* students had to perform all steps correctly, meaning they could not fail the level.

After completing the three training problems, all students completed the Intrinsic Motivation Index (**IMI**) which consisted of 37 questions with sub-scales measuring: interest/enjoyment, perceived competence, effort/importance, pressure/tension, perceived choice, and value/usefulness, (University of Rochester, 2008).

Finally, students took two posttests (counterbalanced with the pretest). In both posttests, all feedback was turned off except for the situational feedback intrinsic to the debate. In **posttest 1**, students had access to the analysis tools they had learned to use during training. In **posttest 2**, students did not have access to the analysis tools.

At the end of the experiment students were asked a variety of demographic questions (not shown in Figure 6.3). Note that students were also given short, 7-question questionnaires on affect after problems 1, 2, 3, 4, and 6, for formative evaluation (not shown in Figure 6.3).

Measures

The first measures in the analyses of learning and interest were:

- **Assistance** (game/tutor). Whether the student was assigned to the *game* or the *cognitive game* version of Policy World. For shorthand, the *cognitive game* version will be referred to as *tutor*.

Policy World automatically records a wide variety of student behavior. This log data was used to construct the following measures for each student on every problem:

- **Time** (min). The time spent on each problem in minutes.
- **Comprehended** (# attempted, # correct). In the first step of analysis, the student had to select text from the report containing a causal claim. Policy World recorded the number of comprehension attempts made by the student, and whether the attempt was correct (contained a causal claim).
- **Evaluated** (# attempted, # correct). In the second step of analysis, the student had to identify the type of evidence of the causal claim, such as *experiment, observational study, case,* or *claim,* and rate the strength of the evidence on a 10 point scale. Policy World recorded the number of evaluation attempts made by the student and whether the attempt was correct. The evaluation was considered correct if: (a) the correct evidence type was specified, and (b) the strength rating roughly observed the following order taught during training: experiments > observational studies > cases > claims.
- **Diagrammed** (# attempted, # valid). In the third step of analysis, the student had to create a diagrammatic representation of the causal claim (if not previously created) and link the claim to the evidence. Policy World recorded the number of diagram citations (number of causal claims linked to an arrow) created by the student and whether the diagram citation was valid. A diagram citation was considered valid if it was linked to a correct causal claim. Note that this is an imperfect measure of diagram correctness. On posttest 1, students were allowed to deviate arbitrarily far from a correct diagram. As mistakes creep into the diagram, it becomes more and more difficult to automatically assess whether a new diagram element is correct. For example, if two contradictory causal claims are linked to the same arrow, it is difficult to automatically

determine if one or both links are incorrect. Validity was used as an approximate (but automated) measure of diagram correctness.

- **Synthesized** (# attempted, # valid). In the fourth step of analysis, the student had to set his overall belief about the causal relationship between the two variables in the new piece of evidence, based on all the previous evidence about those two variables. Policy World recorded the number of synthesis attempts and whether the attempt was valid. A synthesis attempt was considered valid if: (a) the student moved his belief in the direction of the evidence, assuming the student's description of the evidence was correct, and (b) the student's belief mirrored the overall evidence, assuming the student's description of the evidence was correct. Assessing synthesis on posttest 1 automatically suffered from some of the same challenges as assessing diagram construction. Specifically, if the student selected a non-causal claim and synthesized his belief about the evidence, it's difficult to say whether the synthesis is correct. Instead I considered the synthesis attempt valid if it was correct relative to the evidence the student believes he is examining.
- **Recommendation** (# attempted, # correct). During the debate, students had to provide a policy recommendation. Policy World recorded the number of recommendation attempts and whether the attempt was correct. A recommendation was considered correct if the student could win the debate with the given recommendation and all the available evidence.
- **Mechanism** (# attempted, # correct). During the debate, if the student recommended anything except doing nothing, the student would be asked to provide a causal mechanism explaining how the recommendation would affect the desired outcome. Policy World recorded the number of mechanism attempts and whether the attempt was correct. A mechanism was considered correct if it connected the recommendation to the outcome, addressed any relevant issues (variables) and had no causal claims which could be defeated by the opponent (i.e., for which there was more evidence for a contradictory causal relation).
- **Attack** (# attempted, # correct). During the debate, if the student recommended doing nothing, then the opponent would propose a recommendation and mechanism that the student had to attack. The student attacked the opponent mechanism by choosing a causal claim in the mechanism and presenting more evidence for a contradictory causal relation than the opponent could cite for the claim. Policy World recorded the number of attacks attempted and whether the attempt was correct. An attack was considered correct if the attack could succeed given all the available evidence.
- **Evidence** (# attempted, # correct). During the debate, the student had to defend the causal claims in his mechanism or attack by citing more reports in defense of the student's claim than could be cited by the opponent in defense of a contradictory claim. Policy World recorded the number of evidence attempts, and whether the attempt was correct. An attempt was considered correct if the student provided stronger evidence for the claim than the opponent could cite against the claim. Note that to win the debate after providing a mechanism, the student had to win 3 evidence attempts. To win the debate after making an attack, the student had to win 1 evidence attempt.
- **Training success** (0...1). Training success was an aggregate measure of the student's success during training. Training success was calculated by averaging the success rates (# correct / # attempts) of each analysis step (comprehend, evaluate, diagram, synthesize) and debate step (recommendation, mechanism, attack, evidence) on problems 1-3. Training success can also be thought of as a measure of floundering.
- **Debate moves** (qualitative). During the debate, Policy World recorded the student's moves and characters' dialogue including any time the student viewed his diagram.

For each student, Policy World also logged the following sub-scales from the Intrinsic Motivation Inventory (University of Rochester, 2008; Appendix F):

- **Interest** (1-7). The student's interest in playing the game.
- **Competence** (1-7). The student's assessment of how well he played the game.
- **Effort** (1-7). The student's assessment of his effort.
- **Pressure** (1-7). The student's assessment of his anxiety.
- **Choice** (1-7). How much choice the student felt about taking different actions in the game.
- **Value** (1-7). How valuable the student felt it was to play the game to learn about policy.

Results

Analysis 1: Do tutor or game students learn more?

To test whether students learned more from the game or tutor, the first analysis examined success on the posttest 1 analysis steps.

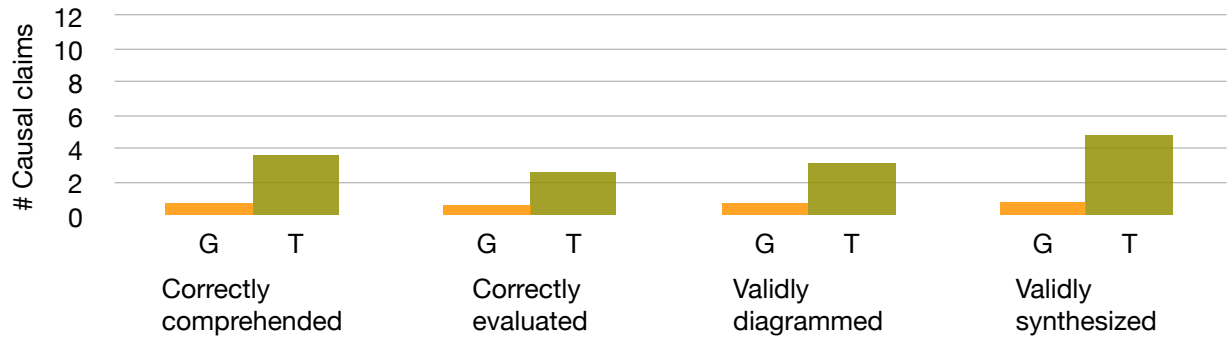


Figure 6.4. Comparison of game and tutor groups on analysis steps during posttest 1.

Table 6.4

Comparison of Game and Tutor Groups on Analysis Steps During Posttest 1

	Game		Tutor					
Measure	**Mean**	**SD**	**Mean**	**SD**	**t**	**p**	**ll**	**ul**
Comprehended	0.80	1.29	3.66	3.22	-9.14	4.1E-18 ***	-5.03	-3.25
Evaluated	0.65	0.95	2.63	2.68	-10.56	1.3E-22 ***	-5.28	-3.62
Diagramed	0.75	1.17	3.16	2.95	-4.69	2.3E-05 ***	-3.44	-1.37
Synthesized	0.85	1.39	4.87	4.69	-5.08	7.7E-06 ***	-5.61	-2.42

Figure 6.4 and Table 6.4 show that tutor students succeed far better than game students on every step of analysis, indicating that the tutor is more effective at teaching analysis. This shows that adding more frequent, knowledge-based feedback, and immediate error correction to a game-based inquiry environment will lead to better learning.

As a second analysis of learning, I analyzed student's performance on the posttest 1 debate. Posttest 1 required students to perform approximately 50 steps. In the analysis phase, analyzing each causal claim required 4 steps (for an average of 48 steps for 12 causal claims). After completing the analysis

phase, students had to debate an opponent which required 3-5 additional steps such as making a recommendation, providing a mechanism, attacking the opponent's alternate mechanism, and providing evidence. The final 3-5 debate steps thus provide a summative measure of students analysis and their ability to use the products of that analysis to make an argument.

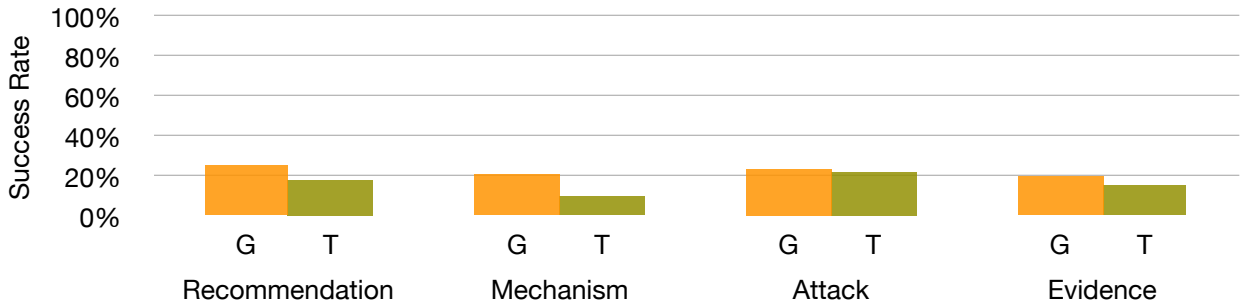


Figure 6.5. Success on each of the four debate steps on posttest 1 for game (G) and tutor (T) students.

Table 6.5

Comparison of Game and Tutor Groups on Debate Steps During Posttest 1

	Game		Tutor					
Measure	**Mean**	**SD**	**Mean**	**SD**	**t**	**p**	**ll**	**ul**
Recommendation	0.26	0.41	0.18	0.28	0.93	0.36	-0.08	0.23
Mechanism	0.21	0.42	0.10	0.31	0.94	0.36	-0.13	0.35
Attack	0.24	0.44	0.22	0.43	0.11	0.91	-0.27	0.30
Evidence	0.20	0.36	0.16	0.31	0.58	0.56	-0.11	0.20

Figure 6.5 and Table 6.5 show the mean success rate (# of correct / # attempted) on each step of the posttest 1 debate for game and tutor students. There was no significant difference between the two groups on any of the debate subtasks. This shows that although tutor students are performing more of the initial steps of problem solving correctly, this advantage does not carry through to the end of problem solving. This result can be understood by analogy to algebra equations. To solve an equation correctly, each step in a long sequence must be performed correctly to reach the correct solution, and even a small number of errors prevents a correct solution. The posttest 1 analysis and debate measures show that tutor students perform more steps correctly, but that both groups are still far from performing the threshold number of correct steps necessary to win the final debate.

Analysis 2: Do diagrams improve argument?

The results of Analysis 1 show a difference in performance between game and tutor students in analysis but not debate. It may be that debate performance does not increase until proficiency in analysis passes a certain threshold. For example, perhaps students' diagrams must be relatively accurate and complete before they can serve as an aid to debate (as in Chapter 3). We can test this interpretation by examining the correlation between the validity of students' diagrams and their debate performance, i.e., we can condition debate performance on diagram validity.

A linear model regressed whether the student won the debate on the number of diagram arrows the student linked to a valid causal claim. Students who created more arrows were more likely to win the debate (b=0.040, t(76)=2.24, p < 0.03). Students who create a greater number of diagram arrows linked to valid causal claims are more likely to win the debate.

The relationship between diagrams and debate performance becomes clearer when we look at how diagramming is associated with the different subtasks of debate.

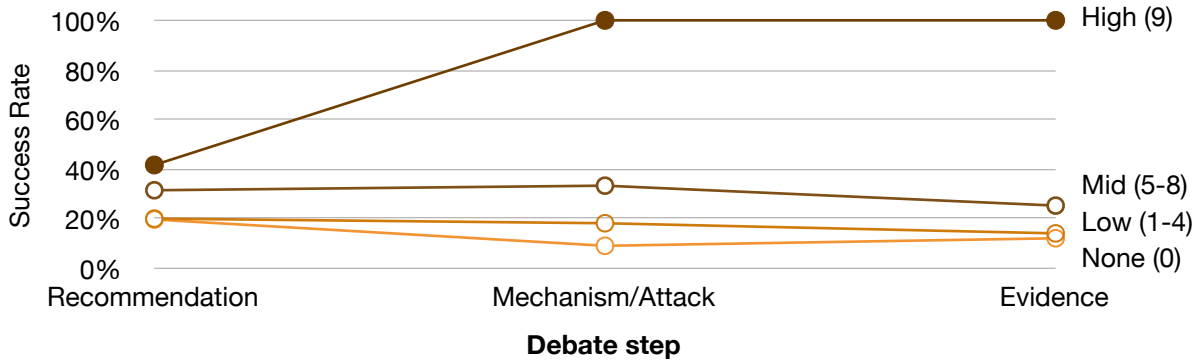


Figure 6.6. Mean success rate on each posttest 1 debate step by size of students' diagrams.

Table 6.6

Mean Success Rate on Each Posttest 1 Debate Step by Size of Students' Diagrams.

	Recommendation		Mechanism		Attack		Evidence	
Valid diagram citations	**Mean**	**n**	**Mean**	**n**	**Mean**	**n**	**Mean**	**n**
9	0.42	3	-	0	1.00	3	1.00	3
8	0.00	1	-	0	0.00	1	0.00	1
7	0.28	3	0.00	1	0.50	2	0.33	3
6	0.25	2	0.50	2	-	0	0.10	2
5	0.50	3	0.33	3	-	0	0.36	3
4	0.17	5	0.00	3	0.00	2	0.00	5
3	0.17	6	0.25	4	0.00	2	0.13	6
2	0.14	7	0.00	4	0.33	3	0.10	7
1	0.26	15	0.20	5	0.30	10	0.21	15
0	0.20	33	0.12	17	0.06	16	0.12	33

Figure 6.6 shows students' success rate on each step of the debate (# correct / # attempts) conditional upon the number of valid diagram arrows they created and linked to valid causal claims on posttest 1. Note that for visual legibility, Figure 6.6 divides students into four groups but displaying each level separately would not substantially change the result, as can be verified in Table 6.6. Figure 6.6 shows that students who diagram substantially more do better during the debate, specifically, they are better able to cite evidence.

Policy World provides no tutoring or feedback for the analysis steps on posttest 1, so the number of claims diagrammed is purely up to the student. Table 6.6 shows that about 40% of students diagram no valid causal claims (0), and another 40% diagram few valid causal claims (1-4). Looking next at

students who diagram between 5-8 causal claims, we see that their performance on the debate is not radically different from those who diagram nothing. What is most striking is the large jump in performance for the 3 students that diagram 9 valid causal claims, especially the jump in providing winning evidence. This may suggest a selection effect, i.e., students who are more diligent both diagram more and do better on the debate. It could also suggest that making relatively incomplete diagrams provides no benefit relative to making no diagram at all. It also seems to suggest that a relatively complete diagram is especially helpful for providing winning evidence.

Analysis 3: Why do diagrams improve use of evidence but not recommendations?

Analysis 2 showed that the number of valid causal claims diagrammed is correlated with providing winning evidence, but what explains this phenomenon? To understand why diagramming is associated with providing winning evidence but not with providing correct recommendations, we can examine several extremes: the high diagrammers who won the debate, the non-diagrammers who won the debate, and the high diagrammers who lost the debate. Specifically: (a) the 3 students who created 9 diagram arrows linked to valid causal claims in the text, all of whom won the debate, (b) the three students who created 7-8 arrows linked to valid causal claims in the text, all of whom lost the debate, and (c) the four students who created no diagram arrows but won the debate.

Note that the following analyses display diagrams reconstructed from students' log data which do not necessarily reflect the visual appearance of the students' actual diagrams. Boxes with thick lines represent variables the student marked as interventions, and ovals with thick lines represent variables the student marked as outcomes. A label like "CC1" indicates that that the arrow is linked to the first valid causal claim in the evidence, "CCX" indicates a clearly incorrect causal claim, and "CC?" indicates the arrow is linked to text related, or in close proximity to a valid causal claim.

Student 116: High-diagrammer who won the posttest 1 debate

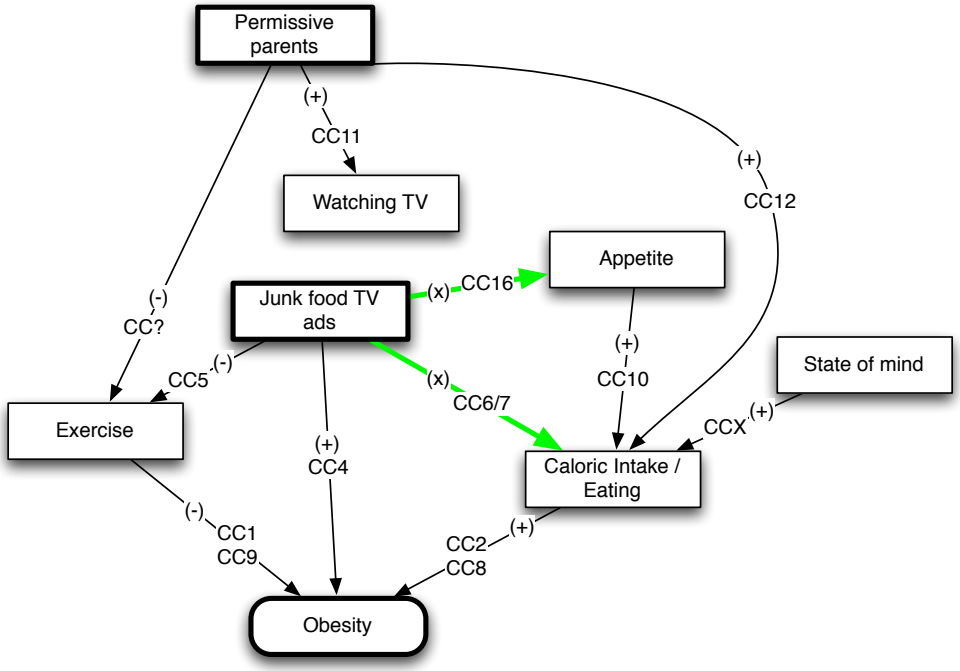


Figure 6.7. The posttest 1 diagram of Student 116 who won the debate and whose diagram contained 9 arrows linked to valid causal claims.

Table 6.7

Debate Transcript for Student 116 a High-Diagrammer Who Won the Posttest 1 Debate.

	Actor	Dialogue	Strike
	Judge	What do you recommend we do about childhood obesity?	
1	*Student*	I think we should decrease parental permissiveness!	!
	Judge	That's not a valid intervention! We don't have control over how parents raise their children.	
	Judge	What do you recommend we do about childhood obesity?	
2	*Student*	I think we should decrease the amount of TV watched!	!
	Judge	That's not a valid intervention! We can't control how much TV people watch.	
		What do you recommend we do about childhood obesity?	
3	*Student*	I think we should increase exercise!	!
	Judge	That's not a valid intervention! Changing the amount people exercise isn't one of the policy options we're considering.	
		What do you recommend we do about childhood obesity?	
4	*Student*	I don't think we should do anything!	
	Judge	According to Harding, decreasing the number of junk food commercials seen will decrease the amount of junk food eaten which will decrease obesity.	
5	*Student*	The number of junk food commercials seen doesn't affect the amount of junk food eaten.	

	Actor	Dialogue	Strike
	Opponent	[Sweating] I don't think there's much reason to believe that the number of junk food commercials seen doesn't affect the amount of junk food eaten.	
6	*Student*	**[Checks diagram 3 times]**	
		[Cites the two reports linked to the arrows in Figure 6.10 marked in green]	
		These 2 reports show that the number of junk food commercials seen will not affect the amount of junk food eaten!	
	Judge	OK, I buy your evidence.	

Table 6.7 shows that Student 116, a high-diagrammer who won the posttest 1 debate, did not check his (or her) diagram when making a recommendation. Even if the student had checked the diagram, it would not have helped, because the student's diagram is incorrect. After making 3 incorrect recommendations, the student recommended *doing nothing* which was the best recommendation in this case given the available evidence. The student correctly attacked the opponent's claim that commercials affect the amount of junk food eaten. The student then checked his diagram 3 times, and finally cited the reports linked to the arrows in green in Figure 6.7, the correct evidence for this attack.

This debate in Table 6.7 shows that the student was able to map the variables in his diagram to the variables being argued in the debate despite the many errors in the diagram. For example, the arrow in the student's diagram from *junk food TV ads* to *appetite* was linked to a causal claim about *junk food ads* and *junk food eaten* (not *appetite*). However, when the student's opponent made a claim about *junk food ads* and *junk food eaten*, this student's diagram was "close enough" that he could look at the report on the arrow from *junk food TV* to *appetite* which was exactly the piece of evidence needed. So the diagram functioned effectively as an index of the evidence even though the errors in the diagram rendered it almost useless for inferring the correct recommendation (which in this case was to do nothing). Of course, the usefulness of the diagram inferring the correct recommendation is irrelevant considering that student did not consult the diagram before making a recommendation.

Student 46: High-diagrammer who won the posttest 1 debate

Student 46, another high-diagrammer who won the posttest 1 debate, was quite similar to Student 116 even though he (or she) began the debate differently. Student 46 began the debate by checking his quite incorrect diagram twice. He then recommended *increasing parental permissiveness*. According to the student's diagram, *parental permissiveness* causally affected obesity but, according to the student's diagram, *increasing* permissiveness would have actually increased obesity which was not the desired outcome. The judge then objected to this recommendation. The student recommended *doing nothing* which was the correct recommendation in this case. The student then successfully attacked the opponent's mechanism. Finally, the student checked his diagram, which had an arrow from *advertising* to *junk food consumption* that was linked to the proper evidence, and cited this evidence to complete a successful attack.

Student 46 used his diagram in a way very similar to student 116. The incorrect diagram, if read correctly, could not have helped Student 46 to reach the correct causal inferences. However the diagram did serve as an effective index of the evidence.

Student 56: High-diagrammer who won the posttest 1 debate

Student 58 was almost identical to Student 46. Student 58 began by consulting his (or her) quite incorrect diagram and proceeded to recommend the implausible intervention of *decreasing parental permissiveness*. After the judge rejected this recommendation, the student then suggested *doing nothing*, which was the best recommendation in this case. Student 58 then attacked the opponent's claim that *junk food advertising increases the amount of junk food eaten*. The student then consulted his diagram and cited the piece of evidence linked to the arrow from *junk food advertising* to *more eating*, successfully supporting the attack.

There is a common pattern among the three high-diagramming winners. They all had initial difficulty finding the correct recommendation, either because they did not check their diagram, or because when they did check their diagram, the diagram had too many errors to help the student infer the correct recommendation. All three students were in a case in which they needed to attack their opponent's mechanism to win, meaning that they did not have to provide a mechanism. They all quickly recognized that they could attack the opponent's claim that *junk food advertising increases the amount of junk food eaten*. Although these students' diagrams contained errors, the diagrams were close enough to the truth that they could find relevant claims in their diagram linked to the needed evidence. Because the diagrams were incorrect, the diagrams could not have helped the students determine the correct recommendation as did the correct diagrams in Chapter 4. But because the diagrams were "close enough", they did help the students find evidence.

Student 29: High-diagrammer who lost the posttest 1 debate

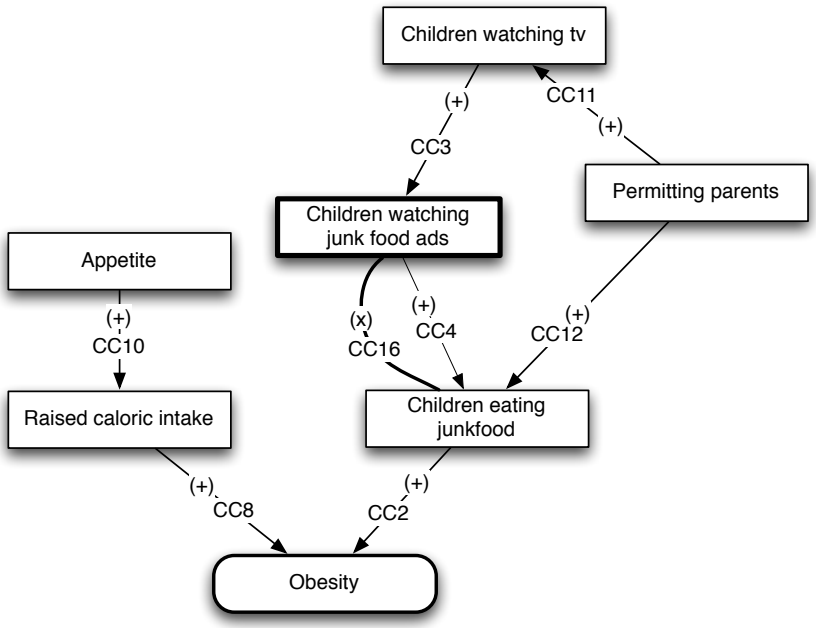


Figure 6.8. The posttest 1 diagram of Student 29 who lost the debate and whose diagram contained 9 arrows linked to valid causal claims. Note that the student's synthesized model stated that *Children watching junk food ads* did not affect *Children eating junk food.*

Table 6.8

Debate Transcript for Student 29, a High-Diagrammer Who Lost the Posttest 1 Debate.

	Actor	Dialogue	Strikes
	Judge	What do you recommend we do about childhood obesity?	
1	*Student*	I think we should decrease the amount of junk food eaten!	!
	Judge	That's not a valid intervention! We don't have a way to directly change how much junk food people eat. What do you recommend we do about childhood obesity?	
2	*Student*	**[Checks diagram]** I think we should decrease parental permissiveness!	!
	Judge	That's not a valid intervention! We don't have control over how parents raise their children. What do you recommend we do about childhood obesity?	
3	*Student*	I think we should decrease the amount of junk food eaten!	!
	Judge	That's not a valid intervention! We don't have a way to directly change how much junk food people eat. What do you recommend we do about childhood obesity?	
4	*Student*	I think we should increase exercise!	!
	Judge	That's not a valid intervention! Changing the amount people exercise isn't one of the policy options we're considering. What do you recommend we do about childhood obesity?	
5	*Student*	I think we should decrease the amount of junk food eaten!	!
	Judge	That's not a valid intervention! We don't have a way to directly change how much junk food people eat. Alright. I've heard enough.	

Student 29 is unusual in that his (or her) diagram (Figure 6.8) was a relatively good representation of the evidence, at least for the causal information needed to win the debate. If the student had interpreted his diagram correctly, it would have led him to a winning recommendation (i.e. do nothing) and mechanism attack (i.e., that junk food advertising has a negligible effect on junk food eaten). Student 29 even checked the diagram after making the first recommendation error. Unfortunately, the student appears not to have interpreted the diagram correctly and went on to not only recommend incorrect interventions (in this case, interventions not under consideration according to the case file), but to repeat those recommendations until he lost the debate.

Student 121: High-diagrammer who lost the posttest 1 debate

The behavior of Student 121, a high-diagrammer who lost the debate, was not radically different from Student 29. Student 121 made 4 incorrect recommendations, and only at the point when there was only one chance left did the student check his (or her) diagram. Student 121's diagram did have the key causal information, but did not label any of the variables as interventions. After checking the diagram, Student 121 then re-recommended his second, incorrect recommendation and lost the debate.

Student 124: High-diagrammer who lost the posttest 1 debate.

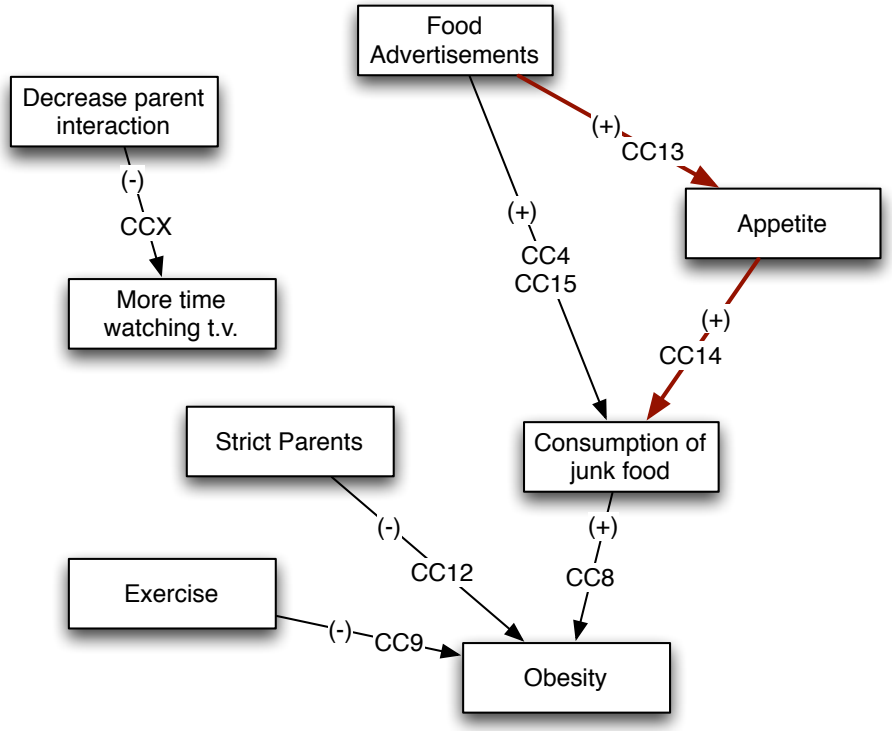


Figure 6.9. The posttest 1 diagram of Student 124 who lost the posttest 1 debate and whose diagram contained 7 arrows linked to valid causal claims.

Table 6.9

Debate Transcript for Student 124 a High-Diagrammer Who Lost the Posttest 1 Debate.

	Actor	Dialogue	Strikes
	Judge	What do you recommend we do about childhood obesity?	
1	*Student*	I think we should decrease the amount of junk food eaten!	
	Judge	That's not a valid intervention! We don't have a way to directly change how much junk food people eat.	!
		What do you recommend we do about childhood obesity?	
2	*Student*	I think we should decrease the number of junk food commercials seen!	
	Judge	How will decreasing the number of junk food commercials seen affect obesity?	
3	*Student*	Decreasing the number of junk food commercials seen will decrease appetite which will decrease obesity!	
	Opponent	There is absolutely no reason to believe that the number of junk food commercials seen increases appetite.	
	Judge	What evidence do you have that increasing the number of junk food commercials seen will increase appetite?	

	Actor	Dialogue	Strikes
	Student	**[Checks evidence linked to arrow from *Food advertisements* to *Consumption of Junk Food* highlighted in red in Figure 6.9. Sees the reports: *Should the Government Regulate Junk Food Advertising?* and *Increasing our Nation's Waistline Through Commercials]***	
4		**[Selects the reports: *Should the Government Regulate Junk Food Advertising?* and *Increasing our Nation's Waistline Through Commercials*]** These 2 reports show that the number of junk food commercials seen will increase appetite!	
	Opponent	That report is irrelevant. [… because the reports are about junk food commericals and junk food eaten, not appetite.]	
	Judge	Yes, I agree. What evidence do you have that increasing the number of junk food commercials seen will increase appetite?	!
	Student	**[Checks evidence linked to arrow from *Food advertisements* to *Appetite* highlighted in red. Sees the reports: *Does Watching Television Increase One's Appetite?*]**	
5		**[Selects the reports: *Does Watching Television Increase One's Appetite?*]** This report shows that the number of junk food commercials seen will increase appetite!	
	Opponent	That report is irrelevant. [… because the report is about TV and appetite, not junk food commercials.]	
	Judge	Yes, I agree. What evidence do you have that increasing the number of junk food commercials seen will increase appetite?	!
	Student	**[Checks diagram but does not view any reports]** **[Checks diagram again looking at evidence linked to arrow from *Appetite* to *Consumption of junk food*. Sees the report: *Increased Exposure to Junk Food Ads Leads to an Increase in Junk Food Ingestion*]**	
6		**[Selects the report: *Increased Exposure to Junk Food Ads Leads to an Increase in Junk Food Ingestion*]** This report shows that the number of junk food commercials seen will increase appetite!	
	Opponent	That report is irrelevant. [… because this report is about junk food commercials and junk food eaten, not appetite.]	
	Judge	Yes, I agree. What evidence do you have that increasing the number of junk food commercials seen will increase appetite?	!
	Student	**[Checks evidence linked to arrow from *Food advertisements* to *Appetite* highlighted in red, then checks the diagram again without viewing any reports]** **[Selects the report: Does Watching Television Increase One's Appetite?]**	
7		This report shows that the number of junk food commercials seen will increase appetite!	
	Opponent	That report is irrelevant. [… because this report is about TV and appetite, not junk food commercials.]	

Actor	Dialogue	Strikes
Judge	Yes, I agree.	!
	Alright. I've heard enough.	

Student 124's debate was quite different from that of Student 29 and 121. Student 124 hit upon the correct recommendation (without consulting his or her faulty diagram) on the second try. The student then proposed an incorrect mechanism, but because the mechanism was plausible (e.g., it began with the recommendation, ended with the outcome, etc.), the judge did not reject it and instead asked for evidence. Then the student repeatedly checked the diagram, but makes two kinds of errors. In this first type error (moves 4 and 6 in Table 6.9) the student checks the wrong arrow in his or her diagram. The judge asks the student about the effect of junk food advertisements on appetite, but in move 4 the student looks at the arrow from junk food advertisements to junk food eaten, and in move 6 looks at the arrow from appetite to junk food eaten. In the second type of error, the student looks at the correct arrow, but the arrow is linked to *irrelevant* evidence. For example, in moves 5 and 7, the student looks at the arrow from junk food advertisements to appetite, but it is linked to text in the report which the expert has coded as a claim about the effects of TV (not junk food advertisements) on appetite.

The high-diagramming students who lost the debate demonstrate some of the ways in which diagramming goes wrong. The student might have: (a) not used the diagram (e.g., Student 121), (b) looked at the wrong arrows or drew incorrect inferences from the arrows (e.g., Student 29, 124), or (c) incorrectly constructed the diagram by linking arrows to evidence that does not support those arrows (Student 124).

Student 127: Non-diagrammer who won the posttest 1 debate

Student 127 made no diagram. After proposing an incorrect recommendation, Student 127 recommended doing nothing which was correct, and then made a successful attack and supported it with winning evidence. It is not clear how Student 127 was able to cite winning evidence.

Student 82: Non-diagrammer who won the posttest 1 debate

Student 82 behaved similarly to Student 127: Student 82 proposed the correct recommendation, provided a mechanism, then successfully provided winning evidence for both of the causal claims in the mechanism.

Student 142: Non-diagrammer who won the posttest 1 debate

Student 142 was slightly different from Student 127 and Student 82. Student 142 began like Student 82 by proposing a correct recommendation and providing a mechanism that could be supported by evidence. However, when defending the mechanism, Student 142 needed three attempts to provide evidence for each of the two causal claims in the mechanism, barely winning the debate.

Student 21: Non-diagrammer who won the posttest 1 debate

Student 21 succeeded by exploiting (probably unintentionally) a loophole in Policy World's debate algorithm in which the opponent will not attack a cause for which the opponent has no evidence.

This meant that if the student created a nonsensical mechanism using a causal claim that did not exist in any report, then the opponent would not attack. Student 21 made use of this weakness by proposing an indefensible mechanism which the opponent neglected to attack at its weakest point.

It is not clear how the 4 students who won the debate without the help of diagram were able to do so. We see only that they made winning debate moves. The first student made no errors, while the second student made only a single error. The third student found a winning recommendation and mechanism quickly, but made 4 errors trying to provide evidence. The students may have recalled the evidence without aid, they may have used pen and paper, they may have inferred the relevant evidence from the titles of the report alone, they may have made a lucky guess, or they may have cheated. The log data do not allow us to distinguish between these possibilities. Note however, that students were not allowed to search for evidence once the debate had started.

If we set aside the fourth student who exploited a loophole and consider the base rate for the 33 students who made no diagrams, we see that only 9% (3/33) won the debate. This suggests that the success of the 3 winning non-diagrammers may just be due to chance. This seems especially plausible when we consider that 3 out of 3 highest diagramming students *all* won the debate. Furthermore, if these non-diagraming students had hit upon a strategy for winning without the need of diagrams, we would expect them to also win the posttest 2 debate which none of them were able to do.

The behavior of the extreme students: high-diagrammers who won, high-diagrammers who lost, and non-diagrammers who won, provides us with several clues as to how diagrams affect debate. We see that even the high-diagrammers did not have diagram construction and interpretation skills sufficiently strong to help them easily provide recommendations and mechanisms. So diagrams are not providing a benefit in the same way as in Chapter 4.

If however, the high-diagramming student managed to stumble upon the correct recommendation and the student's diagram roughly reflected the cited evidence, then the diagram did provide a useful means of indexing the relevant evidence. So much so that the high-diagramming student passed the evidence phase of the debate with relative ease. This explains the huge jump in performance for the evidence step for high-diagrammers in Figure 6.6, and is one of the major drivers of the positive relationship between diagramming and debating. In this sense, the diagram functions as a well organized index.

In short, these analyses of the relations between diagramming and evidence show that students' construction and interpretation skills and dispositions are not sufficient to aid recommendation. Nevertheless, students who create faulty but extensive diagrams demonstrated that they were often able use their diagrams to successfully back claims with evidence.

Analysis 4: Which steps are most difficult to learn?

Analysis 1 suggests that tutor students have greater success with analysis than game students, but that both groups are relatively far from producing strong analyses or debates. On which skills do students flounder, and where is assistance lacking?

During training, game students received minimal feedback and a penalty after making an error in analysis. They were also allowed to end analysis at any point and begin debating their opponent. During training, tutor students received knowledge-based feedback and immediately repaired errors. Tutor students had to analyze all claims before debating their opponent.

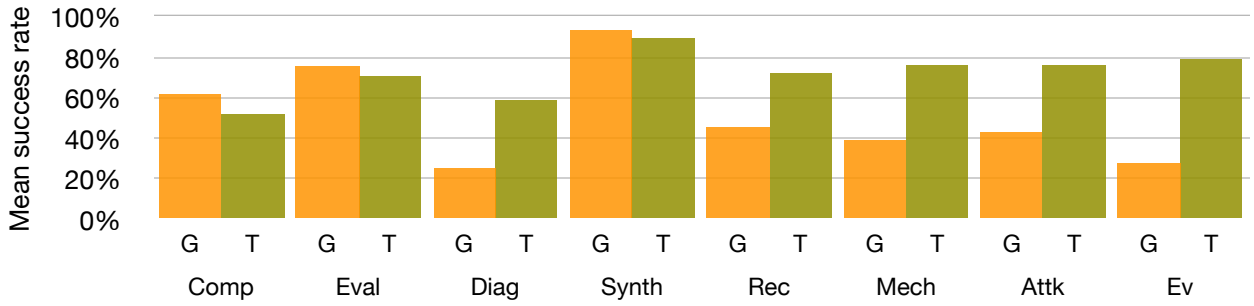


Figure 6.10. Mean success rate on each step of analysis and debate during training problems 1-3 for *G*ame and *T*utor students.

Table 6.10

Mean Success Rate on Each Step of Analysis During Training Problems 1-3

	Game			Tutor		
Measure	**Mean**	**SD**	**n**	**Mean**	**SD**	**n**
Analysis						
Comprehended	0.62	0.32	113	0.52	0.21	114
Evaluated	0.76	0.32	101	0.70	0.17	114
Diagramed	0.25	0.35	91	0.58	0.07	114
Synthesized	0.93	0.15	38	0.89	0.13	114
Debate						
Recommendation	0.45	0.43	196	0.72	0.30	114
Mechanism	0.39	0.48	111	0.76	0.29	82
Attack	0.43	0.49	34	0.76	0.30	32
Evidence	0.28	0.40	171	0.79	0.23	114

Figure 6.10 and Table 6.10 show student's success rate on each analysis and debate step (# correct attempts / # attempts) which tells us how much students flounder on each step during training. These results show that even with assistance, analysis was relatively difficult for students. Table 6.11 shows the mean success rate regressed on step (comprehension, evaluation, diagram construction, synthesis, recommendation, mechanism, attack, and evidence) and type of assistance (game or tutor). The model shows that across both groups, students were more successful on evaluation (2) and synthesis (4) than on comprehension, and less successful on diagramming (3) and debate (5-8). In this model, the overall effect of tutoring was to slightly decrease success (9), for example, the mean comprehension, evaluation and synthesis success scores for tutor students were slightly lower (Table 6.10). This detrimental effect of the tutor was probably the result of a selection effect whereby game students having difficulty with analysis simply skipped the analysis steps (notice the lower n for game students' analysis steps in Table 6.10). In other words, the game selected for higher performing

students on analysis. Despite the slight overall negative effect of the tutor, tutor students were much more successful on diagramming (11) and debate (13-16) than game students.

Table 6.11

Mean Success Rate Regressed on Step and Type of Assistance (Game/Tutor).

	Variable	Estimate (B)	Std error	t	p	
1	Intercept (Comprehension, Game)	0.618	0.03	20.542	1.30E-83	***
2	Evaluation	0.137	0.044	3.133	1.76E-03	**
3	Diagramming	-0.366	0.045	-8.139	7.83E-16	***
4	Synthesis	0.316	0.06	5.274	1.51E-07	***
5	Recommendation	-0.164	0.038	-4.351	1.44E-05	***
6	Mechanism	-0.229	0.043	-5.356	9.72E-08	***
7	Attack	-0.191	0.063	-3.058	2.26E-03	**
8	Evidence	-0.341	0.039	-8.796	3.52E-18	***
9	Tutor	-0.102	0.042	-2.405	1.63E-02	*
10	Evaluation:Tutor interaction	0.05	0.061	0.82	4.12E-01	
11	Diagramming:Tutor interaction	0.435	0.062	7.036	2.91E-12	***
12	Synthesis:Tutor interaction	0.06	0.073	0.815	4.15E-01	
13	Recommendation:Tutor interaction	0.369	0.057	6.499	1.07E-10	***
14	Mechanism:Tutor interaction	0.471	0.063	7.483	1.18E-13	***
15	Attack:Tutor interaction	0.436	0.089	4.875	1.19E-06	***
16	Evidence:Tutor interaction	0.615	0.057	10.708	6.58E-26	***
	R^2	0.282				
	F	42.850			< 2.2e-16	***
	Adjusted R^2	0.275				

These results indicate that the diagram related steps, including diagram construction and diagram interpretation (debate), are the more challenging set of skills, followed by comprehending causal claims. The tutor does seem to be reducing floundering on these steps (Figure 6.10), but we can see that, for diagram construction say, even tutor students construct a diagram element correctly less than 60% of the time. Furthermore, the tutor's assistance does not seem to provide much advantage on the non-diagram steps like comprehension, evaluation, and synthesis.

In addition to recording the number of incorrect attempts, Policy World also collects information about the types of errors made on each step.

Table 6.12

Types on Analysis Errors Made by Students on Evaluation, Diagramming, and Synthesis.

	Error Type	#
Comprehension errors		
1	Quote incorrect	934
Evaluation errors		
2	Evidence type incorrect	286
3	Strength too close to other type	94
4	Quote incorrect	69
5	Strength too high	42
6	Strength violates order	29
7	Strength too low	2
Diagram Errors		
8	Not linked to correct relation	425
9	Variable 2 modifier incorrect	368
10	Variable 1 modifier incorrect	351
11	Inconsistent with links on variable 1	253
12	Inconsistent with links on variable 2	229
13	Inconsistent with links between variables	186
14	Variable 2 already represented	116
15	Variable 1 already represented	73
16	Quote incorrect	60
17	Model has citations linked to invalid quotes	42
Synthesis errors		
18	Belief incorrect	129
19	Belief shift incorrect	67
20	Quote incorrect	56
21	Model has citations linked to invalid quotes	41

Table 6.12 shows the types of errors made on each step of analysis. Note that logs of comprehension errors do not include an error type but rather the text selected by the student which cannot be described in Table 6.12. The table shows that for evaluation, students had the most difficultly identifying the evidence type of the causal claim such as an *experiment, observational study*, *case*, or *claim* (Table 6.12, row 2). Students also made errors determining the strength of the evidence (3, 5, 6, 7). While diagramming claims, students made the greatest number of errors when determining whether a variable was an intervention or an outcome (9, 10), determining which causal claims referred to the same variables (11, 12, 13), and identifying the causal relationship in the evidence (8). On synthesis, students had more trouble with dogmatism, i.e., identifying the causal relation best supported by the evidence (18), than they did moving their beliefs in the correct direction (19). This qualitative data provides a fair amount of detail about students' difficulties that can be used to redesign the assistance provided by Policy World.

Analysis 5: Do tutor or game students spend longer playing?

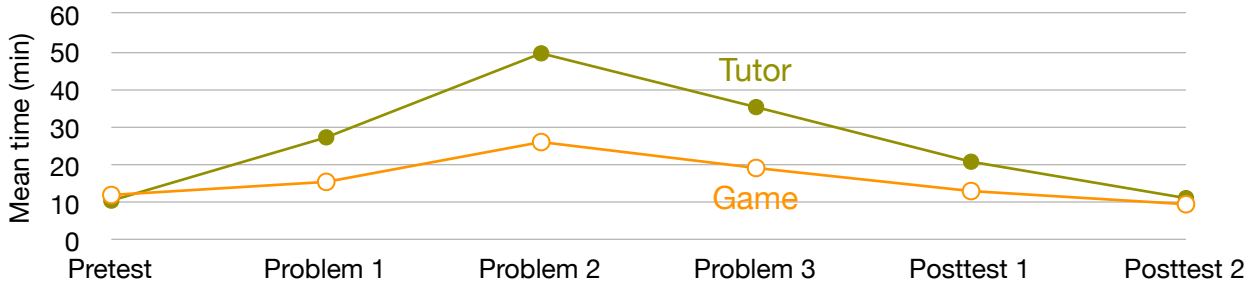


Figure 6.11. Time spent by game and tutor students on each level.

Table 6.13

Time Spent by Game and Tutor Students on Each Level.

	Game		Tutor					
Level	**Mean**	**SD**	**Mean**	**SD**	**t**	**p**	**ll**	**ul**
Pretest	12.09	4.68	10.60	3.76	1.89	0.062 .	-0.08	3.05
Problem 1	15.56	7.91	27.34	15.24	-6.85	1.8E-09 ***	-19.60	-10.77
Problem 2	26.10	17.37	49.58	22.97	-10.00	7.2E-15 ***	-42.20	-28.15
Problem 3	19.24	10.19	35.35	11.71	-11.94	4.1E-17 ***	-28.53	-20.33
Posttest 1	13.13	6.73	20.90	16.72	-2.70	0.01 **	-13.56	-1.99
Posttest 2	9.65	5.94	11.26	8.12	-1.00	0.32	-4.81	1.59

Figure 6.11 and Table 6.13 show that tutor students spent more time playing Policy World than game students. Specifically they spent significantly more time on every training level and on posttest 1.

This raises the question of whether the performance of the game group could be improved by simply increasing time on task. On training problems, game students were automatically promoted after attempting the level twice. It would be relatively easy to simply require students to play the level until they pass the level. If game students show improvement on the second attempt, then increasing time on task should decrease the differences in performance between the two groups.

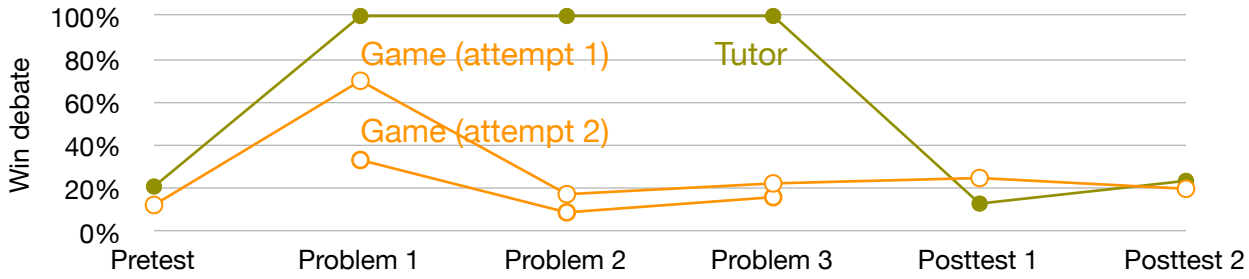


Figure 6.12. Mean number of game and tutor students winning the debate.

Table 6.14

Mean Number of Game and Tutor Students Winning the Debate.

	Game (attempt 1)		Game (attempt 2)		Tutor	
Level	**Mean**	**n**	**Mean**	**n**	**Mean**	**n**
Pretest	0.13	40	-	-	0.21	38
Problem 1	0.70	40	0.33	12	1.00	38
Problem 2	0.18	40	0.09	33	1.00	38
Problem 3	0.23	40	0.16	31	1.00	38
Posttest 1	0.25	40	-	-	0.13	38
Posttest 2	0.20	40	-	-	0.24	38

In fact, Figure 6.12 and Table 6.14 show that game students do not show any improvement on their second attempt. Furthermore, by problem 2, the majority of game students were not passing the level on either the first or the second attempt, despite the fact that the problem does not change between attempts. It seems unlikely that simply asking game students to play longer would significantly increase performance.

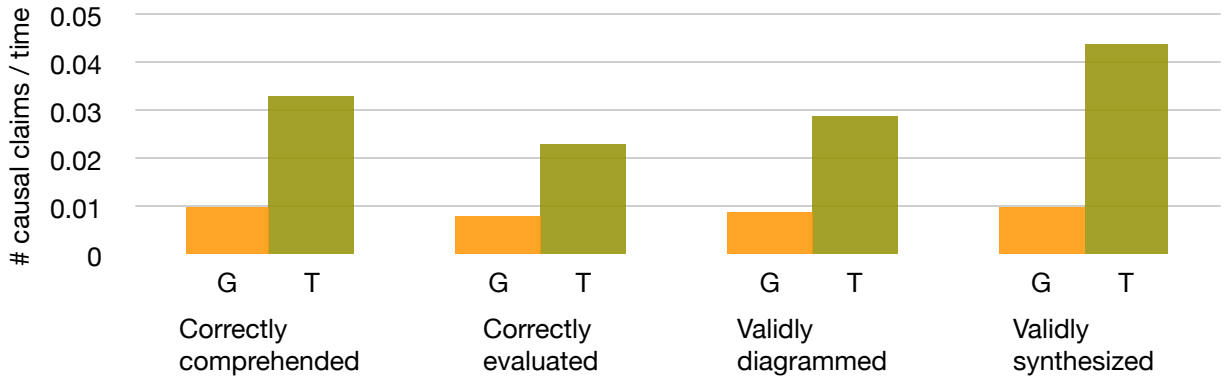


Figure 6.13. Mean number of causal claims successfully analyzed for each step on posttest 1 divided by time on pretest, problem 1, problem 2, and problem 3 for students in game an tutor groups.

Table 6.15.

Mean Success per Minute of Training Regressed on Analysis Step and Type of Assistance

	Game		Tutor						
Measure	Mean	SD	Mean	SD	t	p		ll	ul
Comprehended	0.010	0.015	0.033	0.031	-4.19	1.05E-04	***	-0.034	-0.012
Evaluated	0.008	0.012	0.023	0.025	-3.47	1.06E-03	***	-0.024	-0.006
Diagramed	0.009	0.014	0.029	0.028	-3.78	3.93E-04	***	-0.029	-0.009
Synthesized	0.010	0.016	0.044	0.044	-4.53	4.16E-05	***	-0.049	-0.019

To further investigate the likelihood that the performance of the game group could be improved by increasing time on task we can calculate the learning efficiency of each group. If students learn more quickly in the game condition, then perhaps forcing game students to spend as much time on

training would reduce the differences in learning between the game and tutor groups. Figure 6.13 shows the number of causal claims successfully analyzed on each step of analysis during posttest 1 divided by the number of minutes the student spent on training (including the pretest and problems 1-3). It shows that students in the tutor group learned more quickly than students in the game group. Table 6.15 regressed the number of causal claims successfully analyzed per minute of training time on analysis step (comprehension, evaluation, diagram construction, and synthesis) and type of assistance (game or tutor). The model shows that tutor students learned significantly more efficiently than game students on each step.

This analysis, like the previous analysis, suggests that equalizing the differences in training times between the game and tutor groups will not equalize the differences in learning.

Analysis 6: Do tutor or game students find Policy World more interesting?

One of the primary hypotheses in this study is that a game-based approach to assistance might provide motivational benefits over a tutoring-based approach.

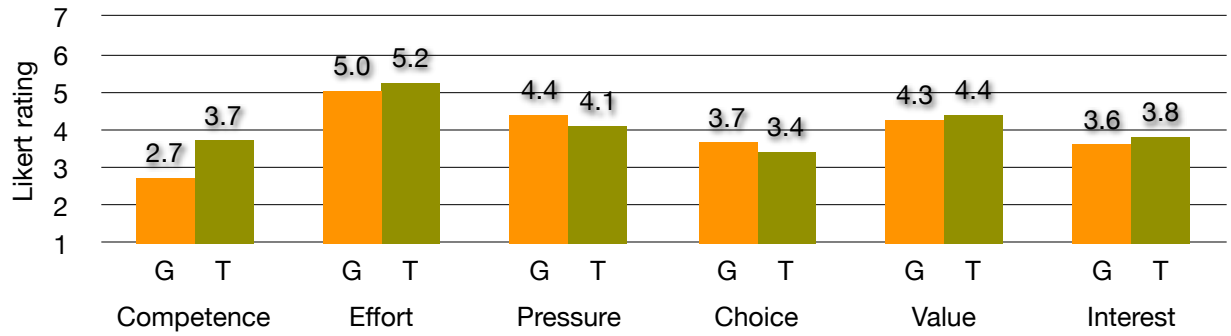


Figure 6.14. Intrinsic Motivation Inventory scores for game and tutor students.

Table 6.16

Intrinsic Motivation Inventory (IMI) Scores for Game and Tutor Students.

	Game		Tutor					
Measure	**Mean**	**SD**	**Mean**	**SD**	**t**	**p**	**ll**	**ul**
Competence	2.72	1.30	3.72	1.25	-3.44	9.54E-04 ***	-1.58	-0.42
Effort	5.03	1.28	5.25	0.84	-0.88	0.38	-0.71	0.27
Pressure	4.39	1.26	4.10	1.22	1.04	0.30	-0.27	0.86
Choice	3.66	1.11	3.41	1.14	0.96	0.34	-0.27	0.76
Value	4.25	1.53	4.40	1.21	-0.46	0.64	-0.77	0.48
Interest	3.61	1.48	3.81	1.19	-0.66	0.51	-0.81	0.41

Figure 6.14 and Table 6.16 show that tutor students report feeling more competent at solving policy problems than do game students. There were no other significant differences on any of the other IMI sub-scales including interest.

Analysis 7: Path Model

Most of the previous analyses used basic statistical tests to examine the relationship between pairs of variables. We can also create path models that take into account background knowledge ignored by these basic statistical tests which allow us to uncover more complex chains of causation among sets of variables. To understand the complex relationships between assistance, training, interest, analysis and debate, I used the GES algorithm implemented in Tetrad 4 (Tetrad 2008; Spirtes, Gylmour & Scheines 2000) to search for equivalence classes of un-confounded causal models consistent with the correlations between the variables (Table 6.17) and prior knowledge about the relationships between variables. This included the prior knowledge that: assistance was determined before any other factor, training was completed next, that intrinsic motivation was measured before posttest 1, that the student created a posttest 1 diagram before debating, and that recommendations were provided before evidence.

Table 6.17

Path Model Correlations for Assistance, Diagramming, Debate, and Motivation.

			Intrinsic Motivation Inventory										
	Assist	Train	Interest	Comp	Effort	Press	Choice	Value	Diag	Rec	Ev	M	SD
Assist	1											0.49	0.50
Interest	.69 ***	1										3.71	1.34
Comp	.08	.19	1									3.21	1.37
Effort	.37 ***	.46 ***	.60 ***	1								5.14	1.08
Press	.10	.14	.19	.05	1							4.25	1.24
Choice	-.12	-.18 .	-.21 .	-.30 ***	.32 ***	1						3.54	1.13
Value	-.11	-.09	.46 ***	.42 ***	-.05	-.16	1					4.32	1.38
Diagr	.05	.09	.80 ***	.55 ***	.19 .	-.10	.36 ***	1				1.94	2.54
Recom	.48 ***	.49 ***	.20 .	.38 ***	.10	-.12	-.01	.22 .	1			0.21	0.34
Eviden	-.08	-.10	-.09	.11	-.14	.03	.05	-.06	.11	1		0.17	0.34
Train	-.05	.05	.10	.27 ***	-.05	-.07	.07	.08	.33 **	.58 ***	1	0.57	0.19

*p<.05 **p<.01 ***p<.001

The best model discovered by Tetrad's GES search algorithm is shown in Figure 6.15. A chi-squared test of the deviance of the path model from the observed values showed that we cannot reject this model a significance level of .05, $\chi 2$ (44, n = 78) = 46.34, p > .38. Note that here, larger p-values indicate better fit and values *above* 0.05 indicate that we cannot reject the model at the a significance level of .05.

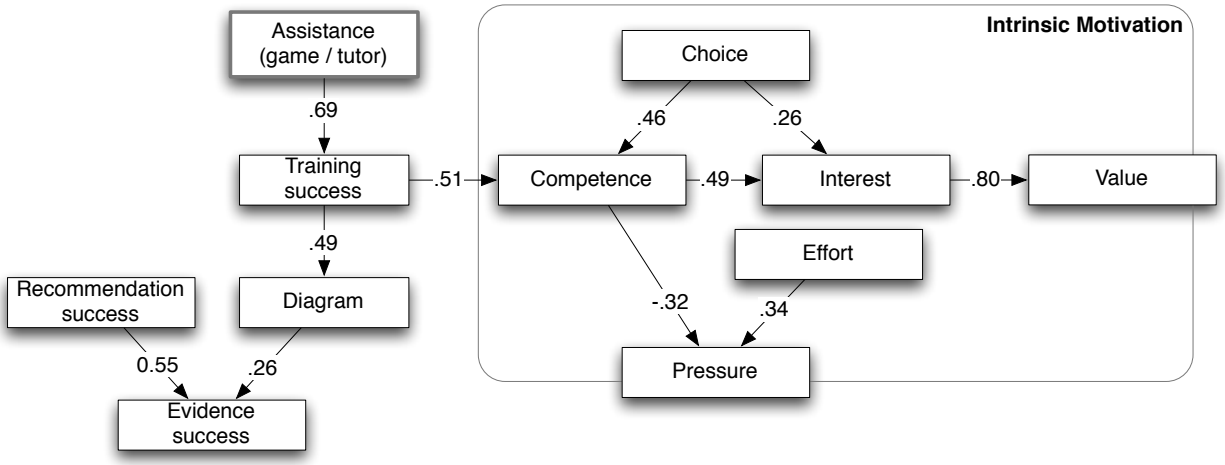


Figure 6.15. A path model analysis of the relations between the assistance provided, success on training, the amount of diagramming on posttest 1, posttest 1 debate performance, and intrinsic motivation.

According to the path model, tutor students had a higher success rate during training (as shown in Analysis 4). Students who had greater success during training were more likely to diagram on posttest 1. Students who diagrammed more were more likely to provide winning evidence (as shown in Analyses 2 & 3). Students who had more success in providing recommendations were more likely to succeed in providing winning evidence, but assistance did not affect student's ability to provide recommendations (consistent with Analysis 3). Those who had greater training success were more likely to report feeling competent (consistent with the correlation between assistance and competence in Analysis 6). Those who reported feeling more competent were more interested in the game, and those with greater interest in the game reported a greater value of the game for learning about policy. Those who felt they had a greater amount of choice while playing the game felt more competent and were more interested in the game. Choice was not affected by assistance.

More generally, the path model elaborates the pathways through which assistance has a beneficial effect on learning to provide evidence. It also shows the positive affect of assistance on competence, interest, and value.

Discussion

The purpose of this study was to: (a) test the effectiveness of combining tutors with games as an instructional approach, and (b) test the effectiveness of Policy World for teaching policy argument. The study showed that adding tutoring-based assistance to a game-like inquiry environment increases both learning and motivation more than using a game-based approach to assistance. The study also showed that Policy World can significantly increase student's ability to reason about policy problems even after a single session, however many learning challenges still remain.

With respect to the first purpose, the study tested three competing hypotheses:

1. **Game hypothesis**: game-based assistance will increase learning and motivation
2. **Cognitive game hypothesis**: tutoring-based assistance will increase learning and motivation

3. **Assistance tradeoff**: game based assistance will increase motivation, while tutor-based assistance will increase learning

Each of these hypotheses corresponds to a different set of claims about the pathways between assistance, learning, and interest. The game hypothesis claims that too much assistance reduces learning and also reduces autonomy and choice which negatively impacts interest. The cognitive game hypothesis claims that assistance reduces floundering and thus increases both learning and interest. The assistance tradeoff argues that the conventional, didactic tutoring will increase learning but decrease choice and thus interest. For the policy reasoning task and game-like inquiry environment used in this study, the results clearly supported the cognitive game hypothesis.

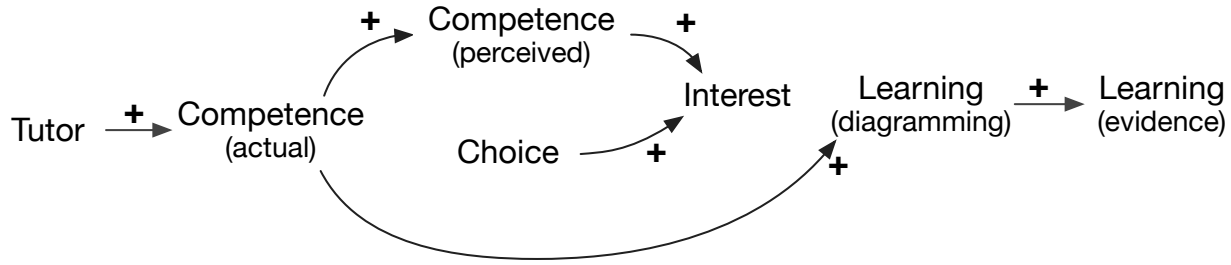


Figure 6.16. Summary of results indicating support for causal mechanisms asserted by the Tutoring Hypothesis.

Figure 6.16 summarizes the results. Adding tutoring to the game-like inquiry environment helped students succeed on training, which increased their ability to create diagrams on the posttest, which increased their ability to cite winning evidence during the policy debate. Adding tutoring also increased students' self-reported confidence which increased their interest in the game (which did not affect learning). Choice *did* increase interest in the activity, however choice was not affected by the tutor. The results can be described intuitively: assistance increased competence which is good for learning and interest. The mechanisms between assistance, learning, and interest described by these results provide consistent support for the cognitive game hypotheses.

With respect to the second goal of increasing students' policy reasoning ability, we saw that Policy World succeeded in improving analysis skills even after 1 training session, and that these analysis skills improve one's ability to provide evidence for policy arguments. More importantly, the log data provided rich information about which skills require greater attention such as diagram construction, diagram interpretation, and comprehension, and which skills are not currently helped by tutoring, such as comprehension, evaluation, synthesis. Furthermore, we now have detailed information about the types of errors that students make. This will allow us to isolate each of these subskills and develop improved tutoring for each step of policy deliberation in future work.

Limitations

One criticism of the study argues that the results would be different if the study had used a better designed game. This is essentially a criticism of environmental validity, because the game-like inquiry environment was held constant across the conditions. This is a fair point. Although Policy World's inquiry environment adhered as closely as possible to the mechanics, dynamics, and aesthetics of *Phoenix Wright*, the Policy World game was quite difficult. While players do not necessarily pass the levels of entertainment games on the first attempt, the performance of students

playing Policy World does seem low. Entertainment game designers face the same dilemma: a game is not fun if it is too difficult, or too easy. When an entertainment game is too difficult, designers respond by making the task easier (Thomson, 2007). Perhaps a better designed version of Policy World would have used easier problems.

This criticism is no doubt true, but misses the point. For the policy reasoning task in this study which had a high, *fixed* level of difficulty, the tutor improved learning and interest more than the game. Making the task easier may have made the game more fun, but it would not have necessarily taught the task we were trying teach. While entertainment game designers can lower the performance bar to make the game more fun, educational games do not necessarily have that luxury. If a tutor increases learning and interest given a task of a certain difficulty, then the conclusion stands. This criticism does point out that the conclusions of this study may be limited to tasks with high levels of difficulty. Future work must establish whether adding tutors to game-like inquiry environments provide the same benefits for easier tasks.

A second related criticism argues that a better designed *educational* game might use different mechanics than traditional entertainment games upon which the game version of Policy World was based. This argument might concede that the *game* version is a accurate representation of how an entertainment game would teach policy reasoning, but argue instead that the *game* version is not an accurate representation of how an *educational* game should teach policy reasoning. I agree. This is precisely the kind of question tested by this study. In this study, the *cognitive game* describes an alternate approach to designing *educational* games by adding traditional cognitive tutors to game-like inquiry environments, and tests this approach against a conventional game-based control.

This criticism suggests that that there are *other* approaches to designing educational games that might rely less on cognitive tutoring. Perhaps the mechanics of entertainment games for children provide a better model. For example, the children's entertainment game *Lego Star Wars* virtually removes penalties. When the child's avatar has lost its *hearts*, the avatar dies by literally falling to pieces and then is immediately regenerated. In a game designed for older players like *Halo*, the traditional penalty for death would be returning the player to some previous checkpoint requiring the player to replay part of the game. When a player dies in *Lego Star Wars*, the avatar also loses some of its *bricks* (which represent money and points), but the player can pick these back up if he is quick enough. At worst, the penalty for death in *Lego Star Wars* is a decrease in score not a salient punishment for preliterate players. A *children's game* approach to educational game design (no penalties, minimal assistance) is halfway between the *game* (penalty, minimal assistance) and *cognitive game* (no penalty, knowledge-based assistance) approaches tested here. Given the results of this study, comparing a *children's game* to a *cognitive game* would be an logical next step for future work.

The study also did not test the effect of the fantasy context on learning and interest. The whole premise behind combining tutors with games is that the game-like inquiry environment increases students' interest in the activity. This is an assumption that should be empirically tested once we overcome the challenges of providing assistance.

The study presented here showed that games and tutors can be productively combined to teach policy argument. While a single session of tutoring is not sufficient to make students expert policy reasoners, the results of the study provide a clear proof of concept for this approach, and its readiness for field-based trials.

7. Conclusion: Towards a curriculum for engaged citizenship

Summary: Democracy depends upon the education of an active engaged citizenry. This dissertation takes a first step toward a scalable, evidence-based civic curriculum by designing and evaluating an intelligent tutor that can teach the skills of deliberation. Specifically it: (a) developed a cognitive framework for deliberation, (b) localized reasoning difficulties in synthesis, (c) showed that causal diagrams can improve reasoning, (d) demonstrated that deliberation can be tutored with computers, and (e) showed that games can better increase learning and interest by using tutor-like assistance. However, even within the limited scope of the tutor, there are many unanswered questions that suggest future work. The first set of questions concerns the learning challenges that must be addressed by the tutor including improving comprehension, diagram construction, and debating skills. Another set of questions concerns instructional strategies used by the tutor including: testing the efficacy of Socratic tutoring, testing the efficacy of the fantasy environment, and isolating the effects of penalties, feedback, and difficulty in game-like environments. A final set of questions concerns increasing the ecological validity of the inquiry environment including adding support for: reframing, searching for evidence via journalistic interviews and scientific experimentation, evaluating bias and scientific information, making arguments about moral values and justice, and the use of persuasion. Fifteen studies are proposed to answer these questions. Completing this multi-year research program would create a solid and expansive evidence-base for the intelligent tutoring of policy deliberation.

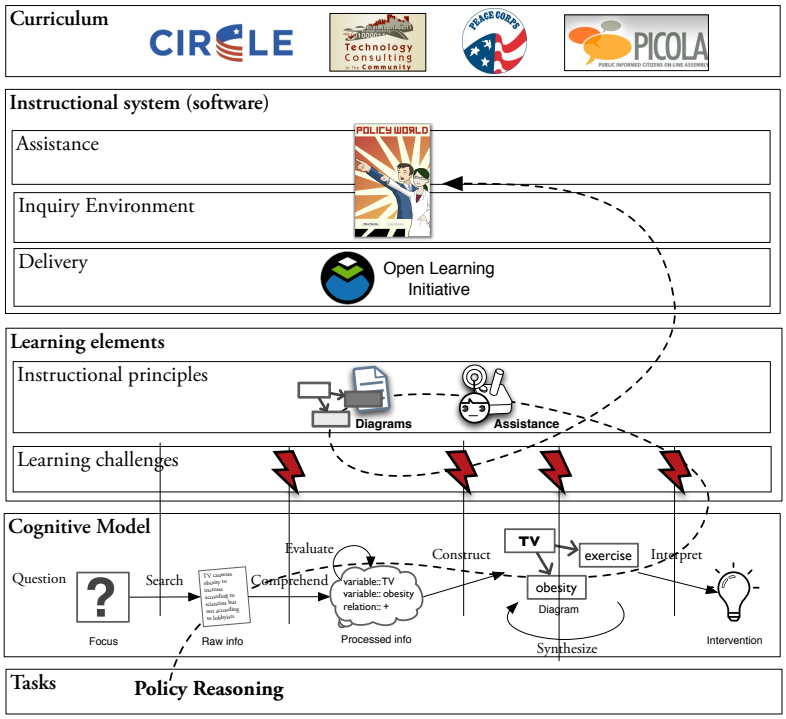


Figure 7.1. This dissertation contributes to multiple levels of the learning environment architecture, and proposes further research on learning challenges, instructional principles, inquiry environment, and assistance.

This work has taken a first step toward the long-term research goal of creating a scalable evidence-based curriculum for civic expertise by designing and evaluating an intelligent tutor that can teach the skills of deliberation, (Figure 7.1).

The results of this work have contributed to our understanding of civic education by providing a cognitive framework for deliberation, showing where errors and bias occur during that deliberation, showing how deliberation can be completed more successfully using diagrams, automated assistance, and game mechanics, and developing an instructional intervention for teaching deliberation. Specifically, the dissertation makes five contributions:

1. **Developed a cognitive framework for deliberation**. Previous research on deliberation proposed models of argumentation that focused on how people reason from recall. These models did not attempt to address how people should incorporate new information, how people make decisions based on evidence, or how to provide automated tutoring. The cognitive framework for deliberation proposed here provides a grammar for thinking and communicating about policy reasoning. The empirical propositions in this work can all be defined in terms of the framework, e.g., where students have difficulties, where the theoretical gaps in previous work lie, and what future goals to focus on. With respect to practice, the cognitive framework for deliberation defines standards for what civic educators should teach. Each chapter of the dissertation showed how the framework concisely describes questions and results and more importantly how it guides instruction.

2. **Localized reasoning difficulties in synthesis.** Previous research predicts that we should see bias in how students *search* for and *evaluate* evidence. However, for the policy tasks used here, the biggest learning challenges, the relatively minor impact evidence has on prior beliefs, seem to occur in the *synthesis* of evidence (Chapter 3). This suggests that research and instruction in deliberation should focus on teaching *synthesis*.

3. **Showed that causal diagrams can improve policy reasoning.** Previous research makes no strong predictions about whether we can effectively use external representations to improve policy reasoning. It also provides little guidance in how to design an effective representation. Research also predicts that constructing diagrams is at best a necessary evil which does not promote learning. However, Chapter 4 showed that not only does providing students with a correct causal diagram improve policy reasoning, but that practicing constructing diagrams helped improve future reasoning when diagrams are unavailable. Chapter 4 also demonstrated some of the serious challenges in learning to construct diagrams. These results suggest the causal diagrams should be used for deliberation, but that we must devote significant effort to understanding how to teach diagram construction.

4. **Demonstrated that deliberation can be tutored.** Some have argued that policy problems are *wicked,* i.e., undefinable, and the intelligent tutoring systems community has had limited success tutoring argumentation or causal reasoning. Probably the single most important contribution of this work is that it shows how to design an instructional system for policy reasoning that provides a level of cognitive feedback approaching that of cognitive tutors for algebra. Chapter 5 demonstrated how a combination of several tutoring strategies: (a) to reify, (b) to limit, (c) to tilt, (d) to use process constraints, and (e) to use student translation, can overcome problems of ill-definition. Chapter 5 also showed how overcoming these challenges of ill-definition along with a simple argument algorithm paves the way for a tutoring system that teaches deliberative argument. Finally Chapter 5 described a pedagogical module that can dynamically switch between direct, cognitive, Socratic, stoic, and game-based modes of feedback. This pedagogical module makes it feasible to combine an intelligent tutor with a game-based inquiry environment, providing a

platform for experimentation. These advances in the tutoring of ill-defined problems open up the possibility of tutoring in a variety of domains where some ill-formed set of information must be transformed into a formal model and used to improve reasoning. Such domains include argumentation, law, history, contextual modeling, and lesson study. Practically, this instructional system shows that we can provide automated tutoring of deliberation, a major step toward the democratization of civic skill.

5. **Showed that games can better increase learning and interest with tutor-like assistance.** There are almost no randomized controlled experiments comparing the effectiveness of tutors and games. Chapter 6 showed that not only can we combine a game environment with a tutoring system, but that using more tutor-like assistance increases both learning and interest. This suggests that educational game designers should consider using intelligent tutoring systems. It also suggests that intelligent tutoring developers can add tutors to game-like environments without substantially altering the way assistance is provided.

Table 7.1

Contributions by Learning Environment Platform Level

Layer	Contribution	Chapter
Curriculum		
Instructional systems		
Assistance	4 Policy World (tutoring architecture)	5
Inquiry environment	4 Policy World (analysis tools and intelligent debaters)	5
Delivery		
Learning elements		
Instructional principles/tactics	5 Tutor	6
	3 Causal diagrams	4
Learning challenges	3 Difficulties using diagrams	4
	2 Location of bias in synthesis	3
Cognitive Models	1 Cognitive framework for deliberation	2
Tasks		

Limitations and future work

The short term agenda for advancing deliberation tutoring research will focus on three broad directions:

(a) **Practical challenges**. Address the usefulness and usability problems with Policy World so that it can be readied for classroom field trials. This will both help us to meet the practical goal of improving civic education, create a platform for future studies, and generate additional research questions that arise from practice. This is the Pasteur's Quadrant, use-inspired basic research approach (Stokes, 1997).

(b) **Learning elements experiments**. Conduct controlled randomized experiments at the learning elements level to evaluate the more important untested design decisions. This will contribute to our basic scientific understanding of how to design game-based learning environments in general, and improve the efficacy of Policy World in specific.
(c) **Ecological validity.** Conduct systems studies to increase the ecological validity of the tasks used in the Policy World. This will expand: the range of learning components that can be taught by Policy World, our knowledge of how to create tutoring systems for ill-defined domains, and create opportunities for additional empirical expert-novice studies of cognition and learning challenges.

Practical challenges

This dissertation provides a proof of concept for use of intelligent tutors for teaching policy deliberation based on evidence from laboratory studies. Now these claims must be tested in classroom-based field research. This will require overcoming a number of practical programming, usability and organizational issues, which while not strictly considered research, must be overcome in order for future research to proceed.

First, there are a number of bugs that must be addressed. Logging must be made more reliable. The OLI delivery environment assumes that Policy World students will complete the exercise in a single session and disconnects them after several hours of inactivity. However, students will often leave their browsers open for a day, and work on Policy World intermittently, in which case their actions may not be logged. A second weakness concerns the overall size of the Policy World interface, which does not display properly on certain screens and browsers. A third type of problem concerns the exploitation of a weakness in the argumentation algorithm. When the student proposes a causal relation that does not appear in any text, the intelligent debaters will not attack the cause, because they do not have any evidence to the contrary. Once these usability issues have been overcome, Policy World's domain model must be adapted to topics suitable for use to meet the needs of particular policy classrooms. Finally, the overall learning curve of cases must be softened to make a more productive learning experience suitable for the classroom. Once these practical matters have been addressed, Policy World can be incorporated into actual public policy classrooms.

Learning elements experiments

Chapter 6 described a test of one design decision in Policy World: to combine tutors with games. However, there are a number of other significant, untested design decisions remaining (a question of instructional principles). There are also a number of deficiencies in the current design (a question of learning challenges). One priority for future work will be to conduct additional experiments like that in Chapter 6 at the instructional principle and learning challenge layers that will test these design decisions and instructional deficiencies.

Future Study 1: Improving diagram construction

Chapter 6 showed diagram construction is one of the most difficult skills for students to learn. The tutor was able to reduce floundering but not enough to prepare students adequately for the posttest. Analysis of the most common error types showed that: (a) students had difficulty identifying types of variables, such as whether the variable was an intervention or outcome, (b) students had difficulty identifying identical variables across causal claims, for example, when one causal claim refers to *junk*

food eaten and another refers to *caloric intake,* and (c) students had difficulty identifying the correct causal relations. Diagram construction can be difficult, because the student must convert a vague verbal description into a formal representation. Alternatively, diagram construction may *appear* to be more difficult than in reality if the tutor's domain model is too rigid, i.e., the tutor may be judging students too strictly.

Consider several examples. When diagramming a variable, the student had to decide whether the variable was a possible intervention, an outcome, or neither. This often required the student to recall information from the policy brief provided at the beginning of the problem which provided information about the possible policy options being considered. It also required the student to make inferences based on common knowledge about what is plausible. For example, if the student wants to decrease *violent behavior* and finds a causal claim that a person's *sex* affects their probability of *violent behavior*, the student should recognize that changing a person's *sex* is not a possible intervention. This example shows how the student must use knowledge not immediately available in the text to take the correct action. This is one way in which diagram construction is difficult.

Consider a second example. When diagramming a causal claim, the student must also decide whether two variables in two different causal claims are the *same*. For example, the tutor's domain model might specify that a claim about *junk food eaten* and a claim about *caloric intake* refer to the same underlying variable. However the student might decide that these two causal claims refer to different variables. At a very fine-grained level this may be true, because these variables might be operationalized in different ways. But if the student uses such a fine-grained interpretation of similarity, he could not relate information about *junk food eaten* to information about *caloric intake*, making it impossible to ever compile enough scientific knowledge to construct a reasonably complete model or to construct models concise enough to guide decision. Unfortunately, Policy World does not explicitly teach at what level of abstraction to consider two variables the same and the domain model does not allow multiple interpretations. In other words, Policy World's domain model may be too strict.

Consider a third example related to *framing*. *Obesity* and *healthy weight* might refer to the same underlying variable measured in the same way. The only difference between the two is that the measure of obesity is approximately the inverse of healthy weight. If Policy World's domain model represents the variable as *obesity*, and the student represents the variable as *healthy weight*, then errors will arise when the student tries to diagram the claim *obesity decreases longevity.* The student may create an arrow from *healthy weight* to *longevity* with a positive causal relation which Policy World will incorrectly consider an error, because its domain model expects a negative causal relation (since obesity *decreases* longevity). The knowledge-based feedback of the Policy World tutor will explain the reasoning behind the assessment to the student, but this example shows some of the subtle ways in which tutoring may be overly brittle.

These possibilities raise several research questions. The first question is whether students are actually making the errors of variable type, variable similarity, and variable framing, described above, or whether the tutor is grading the students' actions too strictly. If tutoring is too brittle, then the second research question is whether we can construct a more intelligent, more flexible tutor. It might be possible to improve tutoring of variable type and variable similarity by isolating the identification of type and similarity step from the diagramming step with increased scaffolding. We could then provide additional instruction and feedback on type and similarity. It might be possible

to improve variable framing by using some limited form of natural language understanding using a domain model with knowledge of synonyms and antonyms to variable names typically used by students. If we find that tutoring is brittle but that we can develop strategies for overcoming this brittleness, then the third research question is whether this improved tutoring actually increases learning or whether brittle tutoring is "good enough".

Future Study 1 asks whether it is possible to define the rules of diagram construction to overcome these problems of ill-definition. A *better-defined* tutor could include a combination of: (a) a more refined expert model with additional rules for representing variable type and similarity, (b) an inquiry environment that uses menus or natural language understanding to identify the content of the students' diagram variables (as opposed to only the structure between variables), and (c) a more flexible pedagogical model that allows multiple representations of variables at different levels of abstraction as well as delayed identification of variable type. Future Study 1 hypothesizes that the *better-defined* tutor will increase our ability to *diagnose diagram construction errors* which will increase *learning* of diagram construction skills relative to the current *process-constraint* tutor. However, the increased sophistication of the *better-defined* tutor may be for naught. The *better-defined* tutor may not increase learning at all, or, by using menus, could actually scaffold past the skill we want to teach. The question is not whether better scaffolding and natural language understanding can improve our ability to diagnose errors in general, but rather whether we can apply these to an ill-defined domain at all.

This study will require three phases. The first phase will require human-coding of the Chapter 6 diagram construction errors to classify which actions are true student errors, and which actions were falsely diagnosed as errors. The second phase will require the development of enhanced diagram construction diagnosis using increased scaffolding and natural language understanding. The third phase will test whether the tutor increases student learning of diagram construction.

The study will consist of a 2-group, randomized, controlled experiment in which one group receives the current *process-constraint* tutor (control), and a second group receives the *better-defined* tutor. In the control group, the *better-defined tutor* will also run silently in order to diagnose diagram construction errors, allowing us to measure the percentage of errors that are incorrectly diagnosed by the current tutor. Pre and post tests on diagram construction will allow us to assess how well each tutor has improved students' learning of diagram construction skills.

The *better-defined* tutor should improve students' diagram construction skills to the extent that it can diagnose (and provide better additional knowledge-based feedback) on diagram construction errors not addressed by the *process-constraint* tutor. Discovering better ways to diagnosis diagram construction errors in policy reasoning will contribute to our basic understanding of intelligent tutoring on the core problems of ill-definition (use of background knowledge and multiple representations) when creating formal diagrammatic representations.

Future Study 2. Improving causal comprehension

Chapter 6 showed that the second most difficult skill for students was finding causal claims in text. Policy World provided relatively little instruction or knowledge-based feedback on comprehension, because we thought that students would have little difficulty with this task based on the results in Chapter 3. Previous work by McCrudden, Schraw, Lehman, and Poliquin (2007) showed that

providing causal diagrams can improve comprehension of causal claims, but here we want students to *create* causal diagrams. We might be able to improve comprehension by teaching students a strategy for recognizing causal claims based on a coding procedure used by Axelrod (1976). The research question asked by Future Study 2 is whether (and how) we can improve causal claim comprehension through a combination of increased direct instruction, scaffolding, and feedback.

Future Study 2 will test whether an *causal schema tutor* can increase students' *learning* of comprehension skills compared to the *error-flagging tutor*. Again, the question is not whether direct instruction and scaffolding work in general, but what specific instruction and scaffolding will improve student's ability to recognize causal claims.

Future Study 2 will consist of three phases. The first phase will examine the comprehension errors in Chapter 6 to determine if there are common patterns of errors or whether the tutor's domain model was too brittle. The second phase will develop an explicit procedure for identifying causal claims based on Axelrod (1976). Axelrod's coding scheme includes a set or rules for recognizing the forms of causal statements that occur in text. The procedure will be validated by analyzing whether or not it would have prevented the Chapter 6 errors if applied correctly. The third phase will test whether the *causal schema tutor* that teaches causal schema recognition rules can improve student learning.

The third phase will use a two-group controlled, randomized experimental design. Students will be assigned to either the *error-flagging tutor* (control) or *causal schema tutor* groups. The *causal schema tutor group* will include more direct instruction on common verbs used to indicate causal claims, such as *increases, causes, leads to,* etc. It might also include better scaffolding of the comprehension step by providing a causal claim *test* with a series of questions such as: *what is the first variable in the claim?, what is the second variable in the claim?, what is the relation between the variables?,* etc.

As in Chapter 6, the experiment will log the number of student attempts and successes at selecting causal claims. It will use a pre/post test design to measure learning and the intrinsic motivation inventory to assess interest. This data will then be analyzed to test the effect of tutoring on learning, floundering, and interest.

The *causal schema tutor* should reduce floundering and thus increase both learning and interest. Tutoring text comprehension is an extremely important skill, and while the rules for detecting causal claims may not transfer to other types of statements, the instructional strategies for teaching these rules should generalize to other domains.

Future studies 1 and 2 seek to address specific learning challenges in policy reasoning that were not taught well by the current tutor. Once we have addressed some of the shortcomings in tutoring, then we can start to test the more general aspects of the tutor. The next set of possible studies will assess the efficacy of some of the design principles for instruction in policy deliberation.

Future Study 3: Difficulty, learning and interest

Chapter 6 discussed the possibility that the *cognitive game* (with tutor) version of Policy World increased learning and interest more than the *game* (no tutor) version only because the problems were so difficult. This criticism argues that a better designed game would have reduced the difficulty level sufficiently so that students would have had a better chance of passing each level. While

lowering the difficulty level may not ultimately be desirable, we would like to know if this would change the results reported in Chapter 6.

Future Study 3 hypothesizes that: (a) there will be no difference in *learning* between the *game* and *cognitive game* versions of Policy World for easier problems, and (b) the *game* version will increase *interest* more than the *cognitive game* version. The first hypothesis reasons that if problems are easy enough to pass with the minimal feedback of the *game*, then it's possible that the knowledge-based feedback of the *cognitive game* will be redundant for easier problems. Of course, this hypothesis may not hold if there are still a significant, albeit smaller, number of students failing the *game* levels who do pass the *cognitive game* levels. The second hypothesis reasons that if students are not floundering in the *game* version, that the extra didactic feedback of the *cognitive game* will decrease student's perceived choice, unlike in Chapter 6.

The experimental setup and design of Future Study 3 differs from Chapter 6 only in the problems provided to students. In this study, the problems will be redesigned so that approximately 80% of students can pass the level in the *game* version of Policy World.

The results of this study will tell us if the benefits of the *cognitive game* approach hold for easier problems. If they do, then we have strengthened the argument for *cognitive game* approach. If they do not, then we must consider the context of instruction when choosing an approach. If students are required to play the game in class or for homework, then we may want to use the *cognitive game* version which improves learning on more difficult tasks. If students have the option to engage in other activities as in an after-school setting, then we will need additional research to determine differences in interest between games with easy problems, and tutors with difficult problems.

Future Study 4. Games vs. children's games vs. cognitive games

Chapter 6 discussed the possibility that the *game* and *cognitive game* constructs are too broad, because they vary multiple sub-variables such as use of penalties, frequency of feedback, and content of feedback. This potential criticism also applies to Future Study 3. Whether or not one accepts the rebuttal in Chapter 6 to this critique, we would like to better isolate the effects of these sub-variables.

Future Study 4 hypothesizes that it is *feedback,* not *penalties,* that increase *learning* and *interest*. This hypothesis argues that the didactic (knowledge-based, step-level) feedback helps students understand the reasons for their errors which they cannot do with only minimal feedback. While penalties may reduce student's tendency to engage in gaming, the penalties do not help students learn.

Future Study 4 will use a two-by-two, randomized, controlled, experimental design that varies the *feedback* [didactic, minimal] and *penalty* [restart, no] imposed by the tutor. This leads the four cells: *game* (minimal feedback, forced restart), *children's game* (minimal feedback, no penalty), *coached game* (didactic feedback, forced restart), *cognitive game* (didactic feedback, no penalty).

The data collection and analysis for this study will be identical to Chapter 6 study except for the explanatory variable. This study will isolate the effects of feedback and penalties, allowing us to make finer grained conclusions than those in Chapter 6. If we find that the *children's game* or *coached game* are superior to the other cells, then the study suggests a different approach to assistance than that provided by either entertainment games or cognitive tutors.

Future Study 5. Socratic vs. prescribed tutoring

Another significant and untested design decision in Policy World was the use of Socratic tutoring during the debate phase of a problem. During the debate, students have to make inferences from their diagrams about what policy to recommend. This requires series of diagram interpretation steps. Rather than ask students to perform each step, Policy World waits for the student to recommend a policy and intervenes only if the answer is incorrect. The Socratic tutor then asks a series of questions isolating sets of diagram interpretation steps. If any of those questions are answered incorrectly, the tutor will then ask another set of questions isolating lower level sets of diagram interpretation steps and so on until the particular incorrect step is found and tutored. The Socratic tutor provides a kind of dynamic scaffolding: more competent students do not have to perform each step, and less competent students only have to perform the subset of steps on which they have difficulty. Policy World uses Socratic tutoring based on the intuition that Socratic tutoring allows us to provide step-level cognitive feedback only when necessary. Future Study 5 will test whether that is an effective decision.

Future Study 5 hypothesizes that *Socratic* tutoring decreases *time* and increases *interest* while maintaining the same level of *learning* relative to *prescribed* tutoring which requires students to perform each step every time. The hypothesis reasons that both *Socratic* and *prescribed* tutoring will increase learning equally, because they both remediate the same errors. *Socratic* tutoring should be more efficient, because it allows students to skip steps on which they are already competent; and because it's more efficient, it should also increase *interest.* Alternately, if we keep time rather than number of attempts constant, the hypothesis predicts that the *Socratic* tutor should increase *learning* while maintaining the same level of *interest* relative to the *prescribed* tutor.

Future Study 5 will use a two-group, controlled randomized experimental design. In this study, students will be given correct diagrams and play the debate phase of Policy World. The study will provide one group of students with the Socratic tutoring of diagram interpretation currently used in the game. The other group of students will answer a question on each step (corresponding to the bottom-out questions of the Socratic tutor) for each debate attempt. Rather than keep the number of problems constant, students will practice debate for a fixed amount of time before playing the final boss battle (posttest).

Future Study 5 will use the same data collection and measures used to analyze debates in Chapter 6 as well as additional logging of the student's responses to the tutoring questions. This will allow us to measure the number of errors remediated by the tutor, and to see whether the *Socratic* and *prescribed* tutor remedy the same number and type of errors and in turn affect learning.

If we keep practice time constant, then the Socratic tutor should focus pedagogical time on the steps on which students have the most difficulty which should increase learning. This may provide the intelligent tutoring community with a new, more effective approach to tutoring.

Future Study 6. Game like inquiry environment

Another significant and untested design decision in Policy World was to use a game-based fantasy environment in order to increase learning and interest. This decision is supported by Cordova & Lepper's 1996 study showing that a fantasy environment increased both learning and interest in a math game. However, more recent studies show no difference in learning between an immersive 3D

fantasy environment and text-heavy environment (Lane, Hays, Auerbach & Core, 2010), so the effects of fantasy environments on learning and interest are by no means clear. The purpose of Future Study 6 will be to assess the effect of Policy World's fantasy context on learning and interest.

Future Study 6 hypothesizes that the fantasy environment will increase both learning and interest, consistent with Cordova and Lepper (1996). This hypothesis reasons that the fantasy environment will increase student's interest which will increase their attention and thus increase learning. Interest did not affect learning in the Chapter 6 study. However we expect that the differences in interest in Future Study 6 to be larger than in Chapter 6 and thus have a more noticeable impact on learning.

Future Study 6 will use a two-group, randomized, controlled experimental design in which one group will receive the current version of Policy World with a game-based *fantasy* environment, and another group will receive a second *no fantasy* version of Policy World with the character dialogue and images removed. All other aspects of the study design will be similar to the study in Chapter 6.

We expect the results of the Future Study 6 to confirm the findings of Cordova and Lepper (1996), showing that the effect of a fantasy generalizes to other domains and types of inquiry environments.

Future Study 7. Boss fights

Another untested design decision concerns Policy World's intelligent debaters. In the current version of Policy World, the intelligent debater is all-knowing. The debater was designed to make the best possible attack based on full knowledge of the domain model and the student's actions. In most games however, the *bosses* become more difficult on each level, which requires the player to become faster, more precise, and more efficient as the game progresses. In other words, boss difficulty provides a kind of scaffolding across levels. In the current version of Policy World, the domain model becomes more complex across problems, but the intelligent debater's difficulty does not. The purpose of Future Study 7 is to test boss difficulty as an instructional strategy for creating a gentler, better scaffolded, learning curve.

Future Study 7 hypothesizes that *boss scaffolding* will increase learning and interest more than *no boss scaffolding* when only minimal feedback is provided. However, when *knowledge-based feedback* is provided, there should be no effect of *boss scaffolding* on *learning* or *interest*. The logic behind this hypothesis is similar to the Future Study 3 hypothesis on difficulty. When only minimal feedback is provided, additional scaffolding will be necessary to make it easier for students to infer the reasons for their errors. When knowledge-based feedback is provided, the scaffolding will become unnecessary.

Future Study 7 will use a 2-by-2, controlled randomized, experimental design that crosses *boss scaffolding* [yes/no], with *feedback* [knowledge-based/minimal], in the presence of *penalties.* Imposing penalties may be necessary in order to encourage students to think about the reasons for their errors rather than just randomly clicking through the boss battles. Because this hypothesis concerns only the debate portion of the game, students will not be required to analyze evidence.

The data collection and analysis will be similar to that in Future Study 5 on Socratic tutoring in that it will use the measures and analysis of the debate phase described in Chapter 6.

The Future Study 7 hypothesis is a variation of that in Future Study 3. Instead of scaffolding the complexity of the domain model, we will also scaffold the difficulty at which the students' skills are tested. This study should provide a further generalization of the effects of scaffolding and difficulty on learning and interest in educational games.

Increasing ecological validity

The policy reasoning task in Chapter 6 had a much larger scope than that in many related studies. Students had to search for information, analyze evidence, and use their analysis to debate. This task and the associated cognitive model advance our understanding of policy reasoning. However, the task and cognitive model are still quite simple when compared to real policy problems. There are a number of ways in which the Policy World task could be expanded to better approximate real policy problems. The following set of proposed studies will describe ways in which the current task falls short of a real policy reasoning task, and how those shortcoming might be addressed. The following studies address the problem of ecological validity, specifically how accurately the task used in Policy World represents real tasks. Few studies in the Intelligent Tutoring Systems/Artificial Intelligence in Education community address ecological validity as the primary research question, so the nature of this type of research question deserves comment.

In many educational domains, ecological validity is a relatively minor issue, because content standards provide the starting point. For example, we do not question whether Cognitive Tutor Algebra should teach equation solving, because equation solving is a learning goal defined by mathematics standards. One *could* question the place of equation-solving skills in mathematical expertise or how these skills transfer to other mathematical tasks, but this does not represent the bulk of ITS/AIED research on mathematics tutors. In the case of policy reasoning and other ill-defined domains which lack established content standards, addressing ecological validity is a serious research challenge.

We can think of a research contribution addressing ecological validity as a kind of *enhanced model* (Newman, 1994), where the model represents a task rather than a technique for predicting performance. For example, a research project for enhancing the ecological validity of a deliberation task by challenging students to *reframe* an issue might be described as follows (based on Newman's 1994 pro forma abstract with filled in slot values italicized):

> Existing *policy deliberation task* models are deficient in providing *framing challenges* in *diagram construction*. An enhanced *policy deliberation task* is described, capable of providing a more realistic *framing challenges* for *diagram construction*. The model has been tested by comparing analyses with empirically measured occurrences of *framing challenges*.

In AIED research, this type of project falls under the category of a *system description*. To demonstrate a contribution, a system description project must include a novel component (a more ecologically valid task) whose benefit is demonstrated through a study of system use (International Artificial Intelligence in Education Society, n.d.). This requires the project to provide evidence that: (a) the relevant attribute of the task is present in the real problems, and (b) the new system teaches some attribute of the task not taught by previous systems. For the following descriptions of proposed studies, I will describe prima facie attributes present in real policy problems that are not currently included in Policy World and describe how these attributes might be included in a future version.

The evidence that will be provided by these research projects is not necessarily derived from controlled, randomized experiments as were the proposed studies on learning elements. For these studies in which the current version of Policy World clearly does not allow certain kinds of tasks, observational data of student's problem solving in an improved version of Policy World should provide sufficient evidence of a contribution.

Future Study 8: Reframing

Framing can refer to a variety of phenomena, but let us consider two specific examples that occur in the news. In a study by Entman, (1991; discussed in Kuypers, 2009), people were provided coverage of a Klu Klux Klan rally which they evaluated more or less favorably if the coverage of the event emphasized free speech as opposed to disruption of public order. The study also compared news coverage of the 1983 downing of flight KAL 007 by the Soviet Union with coverage of the 1988 downing of flight Iran Air 655. They found that when the U.S.S.R was at fault the event was framed as a moral outrage, but when the U.S. was at fault, the event was framed as a technical problem.

These examples of framing can be viewed as emphasizing types of causal antecedents or consequents. For example, one framing of the Klu Klux Klan rally emphasizes the causal effect of the rally on freedom of speech, a desired outcome, while another emphasizes public disruption, an undesired outcome. In the case of the airline downings, one framing states that the downing was caused by an attack, while another framing states that the downing was caused by a technical problem. Although the causes and effects might be established through empirical observations, the differences in framing consist of emphasizing one of two elements in a set of conflicting causes and effects.

In the current version of Policy World, students are given a fixed question, search term, and debate resolution. For example, on the *Cap and Trade* problem, the student's policy brief asks: *Should we implement a cap and trade system for limiting carbon emissions?* Their fixed Google search term is: *cap and trade*, and at the beginning of the debate they are asked: *What should we do about carbon?* In some cases, there may be some minor reframing of the problem. For example in the methamphetamine problem, the student can discover new interventions not mentioned in the policy brief such as decreasing the retail availability of ephedrine which will impact the policy outcome. But for the most part, the student is not allowed to significantly reframe the problem.

An improved version of Policy World would allow students to practice reframing problems in two ways: (a) by reframing the original question by analyzing evidence supporting an alternate frame, and (b) by winning a debate through reframing of the debate question. Policy World can require reframing of the causal antecedent (as in the Klu Klux Klan example) during search and analysis. For example, let's suppose that students begins with a question like: *Should we implement cap and trade*? The initial problem might be framed as Fox News would as a liberal conspiracy to create unnecessary regulation that will ruin the economy. At the beginning of the problem the initial search query would be *cap and trade*. When students find new variables from the domain model in the evidence, these variable names will be added as possible search terms which the student can use to perform additional queries. For example, if the student finds an article about *campaign contributions* and opposition to *cap and trade*, then the student can later search for information about *campaign contributions* which will produce a different set of articles. When the student proposes intervening upon causal antecedents that contradict the original frame, they are essentially reframing the problem. For example, if the student decides that the root problem is not cap and trade, but

campaign finance reform, and if the student proposes that we implement stronger regulations on campaign contributions, the student has essentially reframed the problem in a way that contradicts the original framing of the problem as a liberal conspiracy of cap and trade.

Policy World can require reframing of the causal consequent (as in the Flight KLA 007 example) during the debate. For example, the policy brief might originally frame the problem of cap and trade as a way to prevent environmental crises. During the debate, the student is given a judge that cares little about the environment, but cares a great deal about the economy. In order to persuade the judge, the student will have to show not that cap and trade will save the environment, but that a cap and trade law will spurn U.S. investments in green technology which will help the economy to compete with China in the European renewables market. Changing the argument in this way constitutes a reframing of the original problem in a way that contradicts the original frame.

This improved version of Policy World would allow students to practice one of the most ill-defined aspects of policy reasoning. Additional controlled, randomized experiments can then test learning elements-level hypotheses about how to best provide tutoring of this skill. However, developing an inquiry environment allows this type of task to be a research contribution.

Future Study 9: Search (for journalists)

While most citizens are primarily consumers of policy information via print, television, and internet news, we also want to train some students to be producers of policy information. One way of gathering policy information by journalists and organizers is interviewing. Journalists and organizers interview experts and stakeholders about the causes and effects of a given policy or event, about the source's preferred policy interventions, about the source's desired outcomes, etc. For example, in the Technology Consulting in the Community class, students are taught to interview their community partners about the causal system dynamics within their organizations. Masters of public policy students interview decision makers and stakeholders to identify potentially effective projects. In both classes, students essentially act as journalists to extract policy information at the grassroots community level. In the public policy case, this information is later synthesized with publicly available policy information as found in newspaper articles and scientific reports.

In the current version of Policy World, the information available to students includes only editorials and summaries of scientific studies. Furthermore, the search task is relatively trivial. The search term is fixed, and all the information is within a 2 or 3 click distance from the beginning of the search. Policy World's search environment clearly does not fully represent the search task of interest to public policy educators.

An improved version of Policy World would allow students to search for information by interviewing intelligent agents. One way to create these agents would be to fully specify all the character's possible dialogue in advance. However if the Policy World content author needs to alter the domain model, or the domain model is constructed dynamically to provide evidence that is incongruent with the student's initial beliefs, then content authoring becomes more difficult. Alternately, the intelligent agent could use dialogue automatically generated from a causal model as is currently done to generate the dialogue of the student's avatar. However if there are multiple characters in the game, then there must be some means to differentiate the style of each character's dialogue. It may be possible to use a combination of templates or narrative arcs in combination with causal models to

produce more believable dialogue. No matter how this *intelligent interviewee* is designed, it would then allow the student to ask several types of causal questions, such as: *what are your goals?, what causes x?*, *what does x affect?*, *what do you think we should do about x?,* and *how can we change x*?

The main purpose of this study would be to expand the policy reasoning task to provide a context for students to practice interviewing. It also suggests a number of follow up studies of learning and interest at the learning elements layer. For example, it may be interesting to conduct a controlled comparison of interviewable agents produced by the different approaches described above to see which agents are most believable. We could then test whether believability of agents is associated with increased learning. It also then opens the possibility of additional studies on the more important goal of understanding the obstacles students face while learning to become effective interviewers and how best to teach these skills.

Future Study 10: Search (for scientists).

We also want to train some citizens to produce scientific evidence. In the context of the deliberation framework, producing scientific evidence can be thought of as a type of search. Instead of searching for information that has been produced by journalists, or through interviewing a source, a social scientist searches for data in the environment and analyzes that information to produce causal claims that may be brought to bear on policy debates. For example, a policy analyst may study the European cap and trade system to make a causal claim about the effects of implementing a cap and trade system in the United States.

In the current version of Policy World students are not able to generate scientific evidence.

An improved version of Policy World would allow students to conduct their own observational studies and experiments, as do students using the Causality Lab (Scheines, Easterday, & Danks, 2007). Students would identify variables, select a sample size, and, collect data about the relationship between the two variables by conducting an observational study or experiment. Allowing students to conduct studies in Policy World would require several changes. First, the Policy World domain model would define the *true* causal model with quantitative relations between variables. Second, Policy World would provide an additional search interface for conducting experiments (perhaps modeled after the Causality Lab). Third, a simple natural language generation component would need to be added to produce a report about the results of the student's experiment. At that point, the report generated from the student's experiment could be evaluated as any other piece of evidence in Policy World.

Unlike Causality Lab, the focus in Policy World is for students to select studies that will resolve a policy argument such as conducting experiments to resolve conflicting causal claims. By providing a context in which scientific evidence affects political decisions, adding experimentation to Policy World is a means to show students how science and policy interact.

Future Study 11: Evaluating bias.

In public policy problems, we not only have to worry about the confirmation bias of the citizen, we also have to worry about the bias of the information sources. A large proportion of the information in a public policy debate is provided by sources who are seeking to advance a particular policy agenda. In theory, the professional journalist's and scientist's ideals of objectivity and disinterest

should mitigate this problem, but often it is left to the citizen to assess the bias and integrity of the evidence which requires skills that most citizens lack.

In the current version of Policy World, the causal claims are neutral. Students are not taught to evaluate the source, assess conflicts of interest, and so on. Furthermore, even if students wanted to evaluate the bias and integrity of evidence, the information provided by Policy World does not include the detail necessary for assessment.

An improved version of Policy World would: (a) provide information necessary to assess bias, (b) an interface that reifies the evaluation task, (c) a debate interface that allows students to attack biased information, and (d) an intelligent debater that can assess the student's attack. McManus (2009) provides a *bull detecter* that specifies the kind of questions a citizen should ask when assessing bias and integrity that include 7 types of questions:

1. Who authored the article?
2. Who paid for it?
3. Is the choice of topic socially responsible?
4. Who is likely to be affected by the subject of the story?
5. Does the selection of named sources reveal favoritism or omission?
6. How well are fact-claims supported by evidence?
7. Was there partisan bias in the way the article was framed?
8. Is there ideologic bias or lack of integrity?

The detector defines not only what information needs to be provided by Policy World but how to diagnose student's evaluation. Most of the sub-questions in the detector are specific enough to be phrased as yes/no or multiple choice questions that can be diagnosed by the tutor. The detector helps to define what information the domain model should include for students to assess bias, and how to design an interface that reifies the evaluation task. To use the evaluation of bias during the debate, we need to make two modest changes to the intelligent debater. First, when the debater provides biased evidence for a claim (or counter evidence to the student's claim), the student should have the option to rebut that evidence. Second, if the student does not rebut the evidence, the judging algorithm should accept the opponent's biased evidence. If the student decides to rebut the evidence, then the student will be asked to provide his evaluation of the bias, which if correct, will invalidate the opponent's claim.

Bias is a pervasive feature of policy problems and a key feature of ill-defined domains. An improved version of Policy World that allows the student to evaluate bias and use this evaluation in argument would advance our understanding of how to tutor a core problem cause by ill-definition.

Future Study 12: Evaluating scientific evidence

Policy problems routinely involve scientific evidence. In theory, all citizens are consumers of scientific evidence, and we need to provide them with the tools to evaluate this evidence. However, if we expect citizens to be informed, then they must be more sophisticated evaluators of scientific evidence such as that on global warming.

In the current version of Policy World, the scientific evidence available to students includes only information about whether the study was an observational study or an experiment, which may be a fair reflection of the information available in many newspaper articles.

An improved version of Policy World would allow students to evaluate details of the scientific evidence such as design, measures, sample size, statistic tests, source credibility, and so on. Allowing students to conduct studies in Policy World would require several changes. First, the domain model would have to include the relevant scientific details. Second, the evaluation interface would have to be redesigned in order to allow the student to identify the key pieces of information such as sample size. Finally, the debate interface and intelligent debaters would have to be altered as in *Future Study 11: Evaluating bull* to allow the student to attack scientific evidence based on specific details.

Scientific evidence is an important element in policy debates. An improved version of Policy World that allows students to argue about the specific details of scientific evidence may be a useful tool for training both citizens, social scientists, and policy analysts.

Future Study 13: Values and deontological constraints

Policy problems are often difficult to solve, because they involve conflicts of values. Even if the stakeholders agree on the causal model of the problem, they may still favor different policies. *Future Study 8: Reframing* described how the stakeholders may value different outcomes. A different type of conflict of values concerns which policy interventions are permissible. For example, in the abortion debate, one group considers abortion to be a form of murder while the other does not. This type of restriction on a permissible policy intervention is referred to as a deontological constraint.

In the current version of Policy World, there is no information about sources' deontological constraints, there is no way to represent these constraints in the policy diagram, and there is no way to argue about them.

An improved version of Policy World would allow students to reason about deontological constraints. This would require several changes to Policy World. First, the domain model would have to be changed to specify which stakeholders place constraints on which policy interventions. Optionally, the search interface could also be changed to allow the student to collect polling data about a particular stakeholder's beliefs and knowledge. Second, the diagram construction interface would have to be altered to visually represent these constraints, as well as the corresponding diagram tutoring. Finally, debate algorithm would have to be altered as in *Future Study 8: Reframing* so that the student could make arguments sympathetic to the constraints of the judge.

Note that these alterations to Policy World would only allow the student to recognize the deontological constraints of different sources and argue accordingly. This would not allow the student to argue about which constraints (or desired outcomes) are *just*. Arguments of justice move Policy World away from the realm of deliberation and toward the realm of ethics, although the two cannot truly be separated.

One does not typically use causal models to make arguments about justice, so this would require substantial changes. However, making simple arguments about justice may not be impossible. For example, consider one of the current arguments over a mandate that all able citizens contribute to some health insurance plan. The libertarian position argues that a mandate is unjust, because it

infringes on citizen's liberty. The liberal position (according to John Rawls' difference principle) argues that a just society is one that a rational person would agree to live in even without knowing his original starting position in life. The liberal position argues that no rational person would agree to live in a society in which someone born poor has little chance of receiving proper health care, and that it is just to mandate all citizens to participate in a health insurance fund. One way that philosophers teach justice is to test these moral principles using examples or cases that unsettle our moral intuitions (see Sandel, 2009 for an example). To allow this kind of argument in Policy World, we could allow students to support or attack the permissibility of a deontological constraint by citing a particular case from a list (just like citing a report for a causal claim). Unlike the empirical evidence for causal claims, there is no sense in which Policy World can stack the evidence to make certain arguments about justice "correct" given a particular domain model. So with respect to arguments about justice, the Policy World judge may simply have to be persuaded so long as the student cites an appropriate (if not conclusive) case.

While this improved version of Policy World by no means constitutes a complete ethics tutor, it would allow us to bring issues of justice into the tutoring of policy reasoning in a non-trivial way. In combination with the other proposed studies, this begins to integrate the teaching of policy, science, and ethics.

Future Study 14: Persuasion (via fallacies)

Policy problems are marked by arguments intended to persuade. By definition, advocacy groups that set the policy agenda intend to influence the decision-makers (and thus possibly citizens) who have the power to implement the policy interventions preferred by that group. While some of these arguments appeal to reason, many, if not most, do not. For example, it is not uncommon for partisans to mention the extra-marital affairs of an opponent. Although some may argue that character is important dimension in choosing a decision maker, it certainly is not a valid reason for supporting or opposing a policy intervention such as cap and trade.

Future Studies 8 and 13, describe ways in which the current version of Policy World could be improved to allow reframing and deontological constraints, both of which involve choosing arguments to persuade a given audience on the basis of reason. However, the current version of Policy World does not allow the student to argue about logical fallacies that (unfairly) attack a source's character or that play on the irrational fears of an audience. While we do not want to teach students how to make fallacious attacks, we do need to teach students how to recognize and defend against such attacks.

An improved version of Policy World would require the student to rebut logical fallacies such as ad hominem attacks. It should be relatively easy to allow Policy World to allow rebuttals to logical fallacies. Logical fallacies should not affect the student's policy analysis, so fallacies can be included in the domain model and simply ignored in the student's analysis, or perhaps identified for bonus points, requiring little change in the inquiry environment. For the debate, the intelligent debater could be modified to attack the student using logical fallacies, perhaps when the opponent cannot provide evidence for or against a causal claim. In that case, the student would have to either name the fallacy, or possibly identify one of the limited number of rhetorical strategies for responding to a particular logical fallacy from a list in order to rebut the attack.

In combination with other proposed studies, this begins to integrate the teaching of policy, science, ethics, and rhetoric.

Future Study 15: Complexity

Policy problems like the energy and climate crises can be arbitrarily complex. A full understanding of all the relevant information about climate science, energy production, economics, and policy is beyond the abilities of even the most capable citizen. However, we should assume that citizens can learn to understand much more complex policy problems than of the sort described in this dissertation.

In the current version of Policy World, we see that even domain models limited to 8 variables prove quite difficult for students to solve. Of course students only received a few hours of training, so our expectations should not be over exaggerated.

Each improved version of Policy World should seek to teach students domain models of increasing complexity. At some point, the current policy diagramming tools will not be sufficient for representing these larger problems. Visualizing the causal models at multiple nested levels will become necessary. It is also not clear whether these more complex diagrams will be easily interpreted. Furthermore, when the causal systems contain cycles, it's not clear that the effects of interventions can be predicted without automated assistance or modeling. Finally, constructing these domain models by hand will also become increasingly difficult.

It is at this point that it becomes unclear how to modify the current Policy World system to solve this challenge. We can only see that the challenge must be met. Hopefully, results from the successful completion of the previous 14 studies will provide some guidance.

Table 7.2 summarizes how each level of the learning environment platform is addressed by each proposed study.

Table 7.2

Future Studies by Learning Environment Platform Level

Layer		Future Study
Curriculum		
Instructional systems		
Assistance	1	Improving diagram construction
	2	Improving causal comprehension
Inquiry environment	8	Reframing
	9	Search (for journalists)
	10	Search (for scientists).
	11	Evaluating bias.
	12	Evaluating scientific evidence
	13	Values and deontological constraints
	14	Persuasion (via fallacy)
	15	Complexity
Delivery		
Learning elements		
Instructional principles/tactics	3	Difficulty, learning, and interest
	4	Games vs. children's games vs. cognitive games
	5	Socratic vs. prescribed tutoring
	6	Game like inquiry environment
	7	Boss fights
Learning challenges	1	Improving diagram construction
	2	Improving causal comprehension
Cognitive Models		
Tasks		

Tools

Scaling Policy World for the proposed studies will most likely warrant development of additional analysis and authoring tools.

Tool 1: Problem authoring

Domain models in Policy World are created by hand in XML. However, as the domain models become more sophisticated and the dependencies between pieces of information in the domain model become more complex, the difficulty of debugging the domain models will increase. At some

point, it will become beneficial to create a content authoring tool that partially automates the generation the domain model XML.

Tool 2: Debate visualization

The log data describing students' debate moves and diagrams are analyzed by hand. A tool that can automatically produce debate transcripts or provide a step-by-step debate playback would significantly decrease analysis time.

Tool 3: Automated authoring

Constructing the domain model requires a significant amount of development time. Current work on automated journalism demonstrates that convincing news stories can be created from a combination of quantitative data and domain knowledge. This approach might be used to automatically generate newspaper reports in Policy World's domain model, given a quantitative causal model of the problem, and itself contribute to ITS/AIED research.

Application to other domains

The intelligent tutoring approach used by Policy World can be applied to other domains in which learners must search and analyze information. This includes such disparate domains as human-computer interaction, instructional design, and philosophy. For example, a contextual design tutor for human-computer interaction could allow students to search for key pieces of evidence in video of the user performing some task (comprehension). Students could then use this evidence to construct the five types of user model diagrams specified by contextual design (construction), combine the individual user diagrams into a single representation of the work (synthesis), and finally use the diagram to generate design ideas (interpretation). A lesson study tutor would work in a similar manner. Teachers would watch video of student interviews (search) from which they would identify evidence of students' learning challenges (comprehension). This evidence would then be used to produce a diagrammatic task analysis for each student (construction), summarized across students (synthesis), and used to generate new lesson plans (interpretation). Likewise, a philosophy tutor could be designed in which students read classical texts (search) to identify arguments (comprehension) which they use to produce argument diagrams (construction) that can be combined across authors (synthesis) to resolve a philosophical issue (interpretation). While the types of evidence and knowledge representations are specific to the domain, the basic tutoring approach used here to teach search, analysis, and debate can be used to tutor across a wide variety of ill-defined domains.

Beyond intelligent tutors for deliberation

This dissertation examines only a small part of a larger curriculum on engaged citizenship. Such a curriculum almost certainly cannot be taught entirely through intelligent tutoring. Teaching engaged citizenship will require instruction in several areas such as: deliberation, communication/debate, organizing, and knowledge/experience of the suffering and injustice faced by the disenfranchised (Figure 7.2).

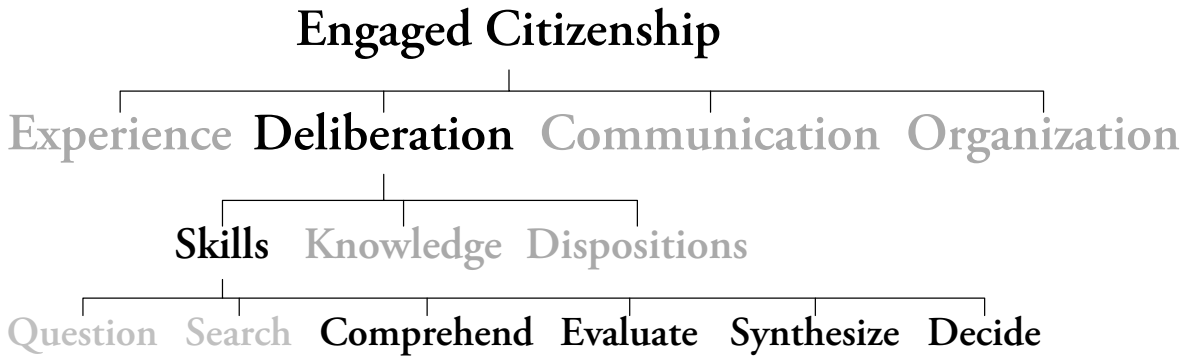


Figure 7.2. The topics of a curriculum for engaged citizenship, only a few of which (in bold) are addressed in this dissertation.

The contributions of this dissertation to this curriculum are modest. The research program for expanding the scope and evidence base for the intelligent tutoring of deliberation is ambitious but achievable. Beyond this work, we need to understand how to teach students how to communicate, discuss, and persuade others about policy, how to work together to act on their knowledge, and how to create opportunities for novice citizens to develop their knowledge, skills, and dispositions.

Appendix A: Inquiry Environments

Table A.1 describes the different types of components used in inquiry environments. **Process** tools scaffold the inquiry process. **Microworlds** (MWorld) provide an opportunity for search and may include fixed data sets (data), simulators (sim), or sensors (sense). **Reference tools** (Ref) allow students to collect annotate, and store information. **Representation tools** (Rep) allow students to create external representations and may include analysis tools (analy) that partially automate the construction process, e.g., creating a bar graph, or modeling tools that partially automate the interpretation process, e.g., by running a simulation of a causal system to show the implications of the representation. While not the focus here, some inquiry environments have other components including **communication** tools that allow collaboration between students and instructors, **delivery** tools that provide access to the inquiry environment, **tutors** that provide feedback as students use the inquiry environment, and **authoring** tools that allow instructors to create content within the inquiry environment.

Table A.1

Different Types of Inquiry Environment Components

	Environment	Domain	Type	Description	Reference
1	Animal Landlord	Animal behavior	MWorld Ref	Students have nine video clips of predator/ prey movies. They can annotate clips using form + drop down list. Annotations are sent to a library (table) with observation & interpretation columns.	Smith & Reiser, 1998
2	Astronomy Village	Astronomy	Process MWorld-data Rep	Students can get assignments from a virtual 2D environment, collect data, and have tools to analyze it.	Dimitrov, Mcgee, & Howard, 2002
3	BioKIDS CyberTracker	Animal Tracking	MWorld-sense Ref	Students use PDA to collect animal sighting data in schoolyard.	Parr, Jones & Songer, 2002
4	Body in Motion	Graphing	MWorld-analy	Students play with motion tracker that graphs data in real time as they move around.	Nemirovsky, Tierney & Wright, 1998
5	Collaboratory Notebook	NA	NA	NA	NA
6	Density Learning Environment	NA	NA	NA	Snir, Smith, & Grosslight, 1995
7	eChem	Chemistry	Rep	Students can construct, visualize, and compare molecules using different views.	Wu, Krajcik & Soloway, 2002
8	Emile	Programming	Process	Programming environment that scaffolds and fades the programming process.	Guzdial, 1994
9	Explanation Constructor	Evolution	Process	Scaffold construction of explanation (seems like the select pre-made explanations and link to evidence).	Reiser, Tabak, Sandoval, Smith, Steinmuller & Leone 2001; Sandoval 2003; Sandoval & Reiser 2004

	Environment	Domain	Type	Description	Reference
10	Galapagos Finches	Evolution	MWorld-data Rep-analy	Construct graphs to compare subsets of populations at different times.	Reiser et al 2001; Tabek, Sandoval, Reiser, Steinmuller, 2000
11	Geometry Tutor	Geometry	Tutor	Classic geometry tutoring.	Anderson, Boyle & Yost 1986
12	Geometer's Sketchpad	Math	Rep-model	Students can construct shapes and equations that can be manipulated/simulated, providing situational feedback like Nathan's ANIMATE.	Jackiw, 1995
13	GenScope / Biologica	Genetics	MWorld-sim	Students can perform genetic experiments and inspect every level, e.g., dna / chromosome / cellular / phenotype / family tree / population.	Horwitz 1996
14	Goal-based Scenario	NA	NA	NA	Schank & Cleary 1995
15	KIE / WISE	Any	Delivery Authoring	Content management and authoring. Create activities with quizzes applets notetaking etc. or use others.	Linn & Slotta 2000
16	Knowledge Forum / CSILE	Any	Rep	Threaded-discussion database where students can create views (diagrams) linking / grouping notes.	Scardamalia & Bereiter 2006
17	Knowledge mediator framework	Evolution	MWorld - data	The evolution thematic investigator contained several elements: (a) conceptual change lesson, (b) hypertext case library that describes five examples of evolution, e.g., the peppered moth, (c) Darwinian model of evolution with five core themes, (d) case commentaries, (e) scientific visualizations, e.g., of Lamarckian evolution, (f) scaffolded problems, (g) guided thematic criss-crossing, i.e., hyperlinks between information resources that could be used to solve synthesis questions.	Jacobson, Sugimoto, & Archodidou, 1996
18	Model-It	Causal	Rep - model	Causal modeling and simulation environment.	Jackson, Stratford, Krajcik, & Soloway, 1996; Metcalf, Krajcik, & Soloway, 2000
19	Media Fusion	Any	Ref Com	Has video clips, e.g. McNeil Leher broadcasts, that the student can annotate with pointers that open up the tabletop data analysis tool which views data also included in the environment. Students can record and send their own video.	Bellamy 1996 (see also tabletop)
20	Personal Assistants for Learning	Newton's Laws	Tutor	Tutor/pedagogical agent, the PAL coaches student in how to solve a problem, then, the student can direct the PAL on how to solve the problem (using menu of directions), and assess how well PAL does.	Reif & Scott, 1999

Environment	Domain	Type	Description	Reference
21 Progress Portfolio	Physics	Ref	Media annotation--the student can store text graphics, sound files, image captures, and annotate; also can use teacher provided templates / and question prompts.	Loh et. al. 1997; Loh, et. al. 2001; Kyza, 2004
22 Project Inquiry	NA	NA	NA	NA
23 Sherlock	Electronics	Tutor	NA	Lesgold, Lajoie, Buno & Eggan 1992
24 SimCalc	Math	Rep - model	Students create functions on a graph and then watch animation of ducks or clowns acting out function.	Roschelle, Kaput & Stroup, 2000
25 Smithtown	Economics	MWorld - sim Rep - analy Tutor	Guided discovery tutor in which the student can make a hypothesis, measure the prices of economic goods, intervene upon the market, observe the resulting changes, and graph the data.	Shute, Glaser 1990
26 SMILE	NA	NA	NA	Koldner, Owensby, Guzdial, 2004
27 STEAMER	Steam engines	MWorld Authoring	A graphical depiction of a navy ship steam propulsion system with tons of views & authoring tool.	Hollan, Hutchins & Weitzman, 1984
28 Symphony	Air pollution	Process Ref	Symphony seems to be a process diagram where students can write a plan, and a file manager for storing their data and graphs. It seems like it's used in conjunction with 4 tools: Artemis (web search), Datawarehouse (collect data), VizIt (graphing), and TheoryBuilder.	Quintana, Eng, Carra, Wu & Soloway, 1999
a ARTEMIS	Air pollution	MWorld - data Ref	Part of Symphony, a background literature search tool -- students can search the UMDL databases, and ARTEMIS stores their links and searches, and allows them to group their materials in questions folders.	Wallace et. al. 1999
b TheoryBuilder	Air pollution	Ref - model Tutor	Part of Symphony, TheoryBuilder is Model It + some tutoring(?)	Jackson, Stratford, Krajcik. & Soloway, 1998
c DataWarehouse	Air pollution	MWorld	Used to collect data on air pollution in Symphony.	
d VizIt	Air pollution	Rep	Used to graph data in Symphony.	
29 TableTop	Data analysis	Rep - analysis	Students can enter data in a spreadsheet, then manipulate data by creating sets, graphs and tables, and also compute descriptive statistics (mean, frequency, cross-tabs etc).	Hancock, Kaput & Goldsmith 1992 (see also media fusion)

Environment	Domain	Type	Description	Reference
30 ThinkerTools	Physics	MWorld - sim	Students engage in instructional cycle of: motivation, model evolution, formalization and transfer using 4 microworlds. In motivation phase, they make predictions about some physics problem. In model evolution, they experiment and record observations in microworld. In formalization, they develop a law to explain observations. In transfer they try to answer the real world question, and if they disagree go back to the microworld for more testing.	White 1984
31 WorldWatcher	Energy transfer in Earth atmosphere	MWorld - data Rep - analy	Visualize global geographic data in color maps. analyze with arithmetic and statistics.	Edelsen & Reiser, 2004; Edelson, Gordon & Pea, 1999
32 Why2-Atlas	Physics	Tutor	Student answers a physical problem in natural language. ATLAS parses the sentence, and converts it to a physics proof; analyzes the proof, and (if there are errors) engages the student in Socratic dialogue to remedy the misconceptions and asks the student to rewrite the essay.	VanLehn, Jordan, Rose, Bhembe, Boettner, Gaydos, et. al., 2002

Appendix B: Representation Tools

Table B.1 lists argumentation, causal, and design-based diagramming tools. A few of the more popular concept mapping, mind-mapping, outlining, and general purpose tools are included to show the range of approaches.

Table B.1

Argumentation, Causal, and Design-based Diagramming Tools

	Tool	Ontology	Description	Reference
1	AquaNet	User defined	Combines NoteCards and gIBIS, but allows the user to define their own types and how those types are spatially arranged.	Marshall, Halasz, Rogers, & Janssen, W. C. 1991
2	Athena	Simplified Toulmin	Argument mapper.	Rolf & Magnusson, 2002
3	ArgMAP	Simplified Toulmin	Argument mapper.	Lau, 2007
4	ArguMed	DEFLog	Argument mapper that uses an extended Toulmin ontology.	Verheij, 1998a;1998b
5	Argutect	Thought tree	Predecessor to Theseus.	NA
6	Araucaria	Simplified Toulmin	Argument mapper.	Rowe, Macagno, Reed, & Walton 2006
7	AVERs	Causal + evidence	Argument mapper.	van den Braak, Vreeswijk, & Prakken, 2007
8	Belvedere	Evidence Maps and Matrices	Collaborative concept mapper and evidence matrix that shows links between claims and supporting data.	Suthers & Hundhausen 2003; Suthers, Weiner, Connelly, & Paolucci 1995
9	Carneades	Toulmin	Toulmin based mathematical model for legal argumentation.	Gordon, 2003
10	ClaimMaker/ Finder/Mapper	Concept map with semiformal ontology for argumentation.	Concept mapping of knowledge claims.	Buckingham Shum, Uren Li, Sereno & Mancini, 2007
11	Compendium	Dialogue map	IBIS mapping tool which is like a concept map with a limited ontology. Nodes can represent issues, ideas, pro, con, and notes.	Buckingham Shum, Selvin, Sierhuis, Conklin, Haley, & Nuseibeh, 2006
12	Convince Me	Evidence map	Creates diagrammatic representations of hypothesis and evidence.	Shank and Ranney 1995
13	CMap Tools	Concept map	Concept mapper.	Cañas, Hill Carff, Suri, Lott, et al. (2004)
14	Debatabase	Simplified Toulmin	Predecessor to Debateapedia.	NA

	Tool	Ontology	Description	Reference
15	DebateGraph DebateMapper	Extended Toulmin	Wiki-based argument mapper. DebateMapper maps can be viewed as a graph, or as a hierarchical list and use an extended ontology. Nodes can be: issues, positions, components, arguments, protagonists, repertoire, scenario, arrows indicate: support, opposition, equivalence, variation, advocacy, relevance, grounding.	NA
16	Debatapedia	Simplified Toulmin	Debatapedia is a wiki containing pro/con arguments for popular debate topics used by the International Debate Education Association. Debatabase is the original pre-wiki version.	NA
17	DIALECTIC	Simplified Toulmin	Argument mapper.	Chryssafidou, 2000
18	Euclid/MacEuclid	Concept map	Concept mapper usually used for argumentation.	Bernstein, 1992; Smolensky, Fox, King, & Lewis, 1988
19	Explanation Constructor	Evidence map	Explanation constructor provides students with causal explanation prompts (arguments) that they can select and then presents a series of prompts that ask students to support the explanation with links to the data they've collected (from the Galapagos Finch environment).	Sandoval 2003
20	Free mind	Concept map	Concept mapper.	NA
21	GARP 3	Causal (qualitative reasoning)	Qualitative reasoning modeler and simulator.	Bredeweg, Bouwer, Jellema, Bertels, Linnebank, & Liem, 2006
22	Genie Software	Causal (Bayesian networks)	Toolkit for creating decision theoretic models.	Druzdzel, 1999
23	Gliffy	Diagram	Diagramming plugin for Confluence wiki.	NA
24	Haystack	Communal notes with user-defined schema	List of searchable notes.	NA
25	HERMES	Zeno (based on IBIS)	Argument mapper.	Karacapilidis and Papadias, 2001
26	HiveLive	Communal notes with user-defined schema	List of searchable notes.	NA
27	iLogos	Simplified Toulmin	Argument mapper.	Easterday, Kanarek & Harell 2009
28	Inspiriation	Concept map	Concept mapper.	NA
29	JANUS	PHI (IBIS)	Capture design rationale in interior architecture.	Fischer, McCall, & Morch, 1989
30	Knowledge Forum/CSILE	Communal notes with explanation templates	Threaded-discussion database where students can create views (diagrams) linking / grouping notes.	Scardamalia & Bereiter, 2006

	Tool	Ontology	Description	Reference
31	LARGO	Legal	Argument mapper.	Pinkwart, Ashley, Aleven, and Lynch, 2007
32	LEGALESE	Legal	Argument mapper.	Hair, 1990
33	Mind Manager	Mind map	Mind mapper.	NA
34	Mind map	Mind map	Mind mapper.	NA
35	Model-It	Causal	Qualitative reasoning modeler and simulator.	Metcalf, Krajcik & Soloway, 2000
36	Media Matrix	Communal notes	Allows students to collect audio and video references off the internet and annotate them for later use in writing.	Kornbluh 2005
37	Notecards	Hypertext	A hypertext editing system sometimes used for argumentation.	VanLehn 1985; Marshal 1987
38	Omnigraffle	Diagram	Diagram tool.	NA
39	OmniOutliner	Mind map	Mind mapper.	NA
40	Oyez	Reference and annotation	Database of supreme court transcripts linked to audio that can be annotated.	NA
41	Philoctopus	Simplified Toulmin	Argument mapper.	Halvatzara, 2007
42	QuestMap	Dialogue map (concept map with ontology: nodes can represent issues, ideas, pro, con, and notes)	Predecessor to Compendium.	Conklin & Begerman 1988
43	Rationale	Simplified Toulmin	Argument mapper.	Van Gelder 2003
44	Reason!Able	Simplified Toulmin	Predecessor to Rationale.	Van Gelder 2002
45	REMAP	Telos (RML, IBIS)	Capture design rationale in software development.	Ramesh & Dhar, 1992
46	Room 5	Legal	Community argument game.	Loui, Norman, Altepeter, Pinkard, Craven, Linsday, & Foltz, 1997
47	Sensemaker	Concept map	Nested concept map used for classroom debate. Predecessor to WISE SAIL/PAS.	Bell 1997
48	Stella	Causal	Students create a quantitative causal diagram which can then be used to simulate a dynamic system.	NA
49	Structured Evidential Argumentation System	Communal explanation template linked to evidence supplied and gathered users.	Helps organize and conduct collaborative argumentation by "intelligence monitors" by looking for evidence on the web or news reports, to fill in argument structure.	Lowrance, Harrison, & Rodriguez 2004
50	Theseus	Thought tree	Argument mapper with a tree of questions and answers that can be used like a simplified Toulmin diagram.	NA
51	Truth Mapper	Simplified Toulmin	A collaborative argument mapping site.	NA

	Tool	Ontology	Description	Reference
52	Visio	Diagram	Diagram tool.	NA
53	VModel	Causal	Students create a qualitative causal diagram which can then be used to simulate a dynamic system.	Forbus 2005
54	Wigmore diagram	Wigmore	A diagram language used for legal reasoning.	Wigmore 1913

Appendix C: Policy games

Table C.1

Games Designed to Address Policy Topics

	Game	Domain	Genre	Description
1	3rd World Farmer	Poor country farming	CMSim	You play a farmer in a poor country. You can plant crops and raise livestock, but instability will wipe out your progress.
2	A Force More Powerful	NA	NA	NA
3	A Seat At The Table	Poverty	Adventure	You play a villager who must make a series of economic decisions, e.g., to join a coffee growing co-op or not. You may survive or meet disaster.
4	Against All Odds	Refugee	Adventure, some action minigames	You play a refugee who must make decisions about how to escape a military takeover, how to survive in a border country and then how to make it as an immigrant.
5	Ars Regendi	NA	NA	NA
6	Ayiti: The cost of life	Public Health	CMSim	You play a rural Haitian family and must make decisions about what work to do, how much to invest in education, health, and community work.
7	Balance of Power	Cold war	Strategy	Cold war simulator
8	Balance of the Planet	Environment	CMSim	NA
9	Budget Hero	U.S. Deficit	CMSim	You are in charge of the U.S. Budget--you set priorities then choose your policies. The simulator then runs for a certain number of years and shows whether you decreased budget and met priorities.
10	Climate Challenge	Policy and climate change	CMSim	You are the president of the E.U. and must reduce emissions while maintaining political popularity. You set national, trade, industry, and domestic policies which have different effects on budget, energy, food, water, and of course CO2. There are also negotiation rounds where you can bribe countries to set emissions targets.
11	Community Organizing Toolkit	Community organizing	Adventure	You play a community volunteer. You get training on how to conduct a door knocking campaign, e.g., what houses to approach, how to introduce self, how to build rapport, how to get commitment.
12	Connect 2 Climate	Environment	Action	You play cell-phone action games such as flying a plane to avoid obstacles. If you make an error, you get a fact about energy use and climate.
13	CONSENT!	NA	NA	NA
14	Darfur is Dying	Genocide	CMSim & Action	You are a refugee in Darfur. You play an action game to get water and avoid Janjaweed, then distribute the water in the camp to build housing and grow food. There are random attacks on the camp which destroy the housing. After a raid, the player is prompted to take action via the Darfur is Dying website.
15	Deliver the Net	Mosquito nets	Action	You drive a motorcycle through the countryside to deliver nets to people. After playing, you are prompted to make a donation.

Game	Domain	Genre	Description
16 eLECTIONS: Your Adventure in Politic	Presidential elections	Adventure	You run a campaign for the presidency against opponent. You choose a platform at the beginning of the game and your main choices are: (a) what states to campaign in, (b) what fundraisers to do (you need to choose fundraisers that support your position), and (c) how to respond to scandals.
17 ElectroCity	Energy	CMSim	You manage the energy resources of a New Zealand town. You have to balance energy management, political popularity, population growth, and environment by develop the land and energy resources and buying energy on market.
18 Energyville	Energy	CMSim	Play two rounds of developing energy resources for a city in this Chevron sponsored game.
19 Fat world	Obesity	CMSim	You are a resident of Fatworld. You can buy a house, plan and shop for meals, buy a restaurant to make money, exercise, and buy political influence.
20 Fantasy Congress	NA	NA	NA
21 Food Force	Food aid	Action mini-games embedded in story	You play a world food program worker. The game presents a story with video clips and description of parts of food aid process, e.g., survey crises, develop nutritional mix of food, initial food drop, purchase food, deliver food, invest in farming and development. For each part of the process you play minigame where you carry out the process.
22 Free Rice	Elementary math, vocabulary	Human computing	You solve math problems or define words, and every correct answer donates 10 grains of rice.
23 Global Conflicts	Latin American / Palestinian politics	Adventure with 3D engine	You play a journalist investigating problems like the Macquiladoras. In the first phase, you search for information by interviewing people, which gives you arguments and statements. You then use these to cross-examine the villain, e.g. the factory boss who will crack if you ask the right questions and use the right argument at the right time.
24 Go Goat Go	Poverty & livestock	Action	You play a boy in a Kenyan villiage that has received a goat from Christian Aid. The story explains how goats provide milk, and dung for fertilizer, and during the game you milk the goat, collect dung and spread fertilizer.
25 GumBeat	NA	NA	NA
26 Harpooned	Whaling	Action	You play a Japanese "scientific research" whaling vessel that hunts whales and turns the meat into food using a Spy Hunter like mechanic. You have to avoid harpooning the protestors. Eventually, you will be killed by the vigilante protester boat.
27 Heifer Virtual Village: Nepal	Poverty & livestock	Adventure with 3D engine	You play a girl in a Nepalese village. Your mother gives you some clear cut tasks that involve walking back and forth between different locations to collect firewood, recruit people for a community center to raise money, build a goat pen, and buy a goat. The story/dialogue emphasizes the importance of getting a goat for economic development.

Game	Domain	Genre	Description
28 Hidden Agenda	Global politics	CMSim	Government sim. You play the president of Chimerica (a fictional Latin American country).
29 Homeless: it's no game	Homelessness	Adventure	You walk around a maze and try to collect money and bottles, and use the bathroom. You can be stopped by the police who will take your things.
30 Hush	Rawandan genocide	Rhythm	You are a Rawandan mother who must sing her child to sleep as the village is being attacked. As you tap out the letters, an cinematic montage unfolds in the background.
31 ICED - I Can End Deportation	NA	NA	NA
32 Karma Tycoon	Non-profit management	CMSim	You play a non-profit manager who applies for grants and opens centers (e.g., youth centers, homeless shelters etc.) The more services you provide the more grants you can apply for.
33 Layoff	2008 Economic crisis	Puzzle	You play a Tetris like game where you move around workers to lay them off. Managers cannot be laid off. If you get stuck, you can click the bank bailout button to shuffle the pieces.
34 LegSim	NA	NA	NA
35 Mission: Migration	Habitat destruction	Action	You play a migrating flock of birds. In the first minigame, you fly over rural, suburban, or city and must avoid obstacles like clouds and airplanes. In the second mini-game, you have to land your flock on the ground while avoiding hazards like pesticides and aiming for bonuses like bird feeders.
36 My US Rep	Politics	Action-adventure	You are a U.S. representative walking around trying to pick up cash while avoiding protesters and lobbyists (literally). When you touch a bill, you see the rep's actual vote and say whether you agree or disagree. Depending on what you chose, the popularity of the rep goes up or down. If you vote too much with the party, more lobbyists show up on the board, if you get too unpopular, you lose.
37 Nation States	Politics	CMSim	You play the government of a fictional nation. Each day, you vote on one issue which affects your nation's civil rights, economy, and political freedoms. You can also join a world assembly and submit resolutions that are voted on and binding.
38 Nuclear Weapons: The Peace Dove Game	Nuclear-weapons	Quiz	Jeopardy style game where you are given a description of a nuclear power and given 2 chances to identify the country.
39 Oiligarchy	Oil & politics	CMSim	You play an oil company which can explore for oil, drill wells, and bribe politicians to protect or open up new oil resources. If you succeed, the world runs out of oil and blows itself up; if you fail you get fired.

Game	Domain	Genre	Description
40 Operation: Resilient Planet	Ecology	Action-adventure	In mission 2, you play an Argonaut, who needs to research endangered turtles to assess environmental impact of decommisioning an oil rig. You complete a series of missions: (a) take pictures of animals, classify and identify animals, and sort animals into food web, (b) search for habitat, threat, diet, ecosystem role of two different species of turtles, (c) make arguments about the turtles by picking among 2 claims, then using the observational data from (b) to support the claims, and finally, (d) tag the turtles to delay the oil rig detonation.
41 Orange Revolution	Ukrainian Orange Revoluation	Adventure	You play both candidates in the 2004 election -- depending of the decision of both candidates, the country either descends into riots, or concludes the election peacefully.
42 Our courts	Rule of law	Action-adventure	Currently in development, you talk to different characters to collect argument cards which you use to win court cases.
43 PeaceMaker	Israel-Palestine conflict	CMSim	You play either an Israeli or Palestinian leader. The Israeli leader can take security, political, or developmental actions and must balance their effects on different interest groups.
44 Pictures for Truth	NA	NA	NA
45 Play The News	News	Quiz	You: (a) see a current news issue, e.g., bill to ban fast food franchises around depressed neighborhoods, (b) can read background information and stakeholder stances on issue, (c) choose to play as one of the interested parties, (d) say what you think should happen and see what other people have said, and (e) predict what actually will happen. You can win the game if the interested parties act according to your prediction.
46 Pos or not	HIV	Quiz	You are shown a picture of a person and guess whether they are HIV+ or not, then results tell you if you were right.
47 Raid Gaza	Israel-Palestine conflict	CMSim	You play the Israeli military, you can build military installations that allow you to attack with different weapons. You can also call the US for more funds at any time. At the end of the game, if you achieve a lower than 25:1 casualty ratio (the real ratio) you lose.
48 Real Lives 2010	NA	NA	NA
49 RePlay: Finding Zoe	Abuse and stereotypes	Adventure with action mini-game	You play teenage friends of Zoe who seems to be in an abusive relationship. You go around town collecting pages of her journal and talking with other kids who say negative stereotypes about Zoe to which you respond by dragging words onto a thought bubble to match either a put down or denial of the stereotype -- if you say the right response, the kids join you. If you push the kids out of the way, you get bad karma and can't find Zoe until you apologize. At the end you all go to Zoe and show her how many people like her. Game ends with strange trampoline/human pyramid mini-game.

Game	Domain	Genre	Description
50 RoboRush	Small business management	Adventure	You sell robots. You have to: (a) go door to door, (b) find out what the customer wants, or something else they're willing to buy that may be more profitable, (c) construct the robot. On level two, you open a shop and hire assistants (whom you must train). On level three, you build a factory and specialize in a particular kind of robot -- you do well if you pick a robot type that is both profitable and popular.
51 September 12th	War on Terror	CMSim	You see a village with people and terrorists walking around. You can fire a missile to kill a terrorist, but doing so will create more terrorists than you destroy.
52 Serious Policy	NA	NA	NA
53 SimEP (Solid Waste Management Computer Game)	Solid waste management	Adventure with action mini-game	You play a citizen of Hong Kong. You must gather 50 stars which you earn by reading newspaper articles and commercials about waste management policy, and by taking actions to reduce waste such as recycling. When do something good, the amount of time Hong Kong has left until the landfills are full increases; if you do something wrong, it increases. You also play some action-mini games such as sorting the trash.
54 Stop Disasters	Natural disasters	CMSim	You prepare and develop an area for natural disasters such as tsunamis, floods, and earthquakes. You can build emergency warning systems, reinforce buildings, and create protective landscaping.
55 The Arcade Wire: Airport Security	Airport security	CMSim	You play an airport security guard. Constantly changing alert messages tell you what items are currently restricted/ allowed, e.g., shirts, snakes, etc. You can make 3 errors of commission/omission before you lose.
56 The Arcade Wire: Oil God	Oil & politics	CMSim	You play an "Oil God". You can inflict wars, natural disasters, change economies etc. with the goal of doubling the cost of gas. Disrupting the flow of crude increases the price.
57 The Budget Maze	City budget process	Adventure	You navigate through a maze in which you have to make decisions about who to talk to/what organizing actions to take in order to fund a certain program. If you take too long or don't acquire enough clout, your program doesn't get funded (note, some levels don't seem possible to win).
58 The Garbage Game	Solid waste management	Adventure + CMSim	Quizes you about how you would handle your personal waste and cities waste, then tallies the economic and environmental impact.
59 The Great Green Game	Environment	Quiz	You are asked multiple choice questions mostly about the effect of consumer choices on energy and the environment.
60 The Redistricting Game	Redistricting	Puzzle	You play a politician in charge or drawing the district map -- you try to satisfy the redistricting rules while maintaining party majority in each district. Different people will approve/disapprove of the plan, and the opposing party will challenge the map on various grounds in court.

Game	Domain	Genre	Description
61 The Vinyl Game	Plastics and environment	CMSim	You manage the vinyl production process including production, compounding, manufacturing, use, and waste management. You try to maximize production and environmental sustainability.
62 Traces of Hope	NA	NA	NA
63 World Without Oil	Energy	Alternate reality	World Without Oil was a website that imagined a 32 weeks of the oil shock where supply no longer meets demand. Each week, the website posted some fictional events, e.g., price of oil goes up, food crises, riots, etc. and then all the players submitted an expression of what they imagined their lives would be like in the form of blogs, videos, images, comics, real-life actions such as protests, etc.

Appendix D. Argument moves

The following table describes the different argument moves made during the debate portion of the game as described in Figure 5.7. Each move represents a semantically different argumentation move object, but not how that move is translated into text and presented to the student – that translation is done by a separate software component that may present the move differently depending on the current problem/context of the game and often will translate the move into multiple speech acts. Note that the activity names preceded by the role of the character making the move, e.g., Player, Judge, or Opponent, but different roles may be played by different characters, e.g., during a training problem, the Judge and Opponent roles might both be played by the mentor character. Note also that Player refers to the game character that represents the student.

Table D.1

Argument Moves Used by the Intelligent Debaters in Policy World

Activity / Move	**Explanation**
1. Judge: Ask recommendation	
Ask For Recommendation	The judge asks the student to make a policy recommendation, e.g., "What should we do to decrease childhood obesity?"
2. Student: Pick recommendation	
Pick Recommendation	The student picks a recommendation from a list of options including increasing/decreasing each variable in the model or doing nothing, e.g., "decreasing junk food advertising"
3. Player: Say recommendation	
Say Recommendation	The player character states the policy intervention picked by the student in prose, e.g., "I think we should decrease junk food advertising."
4. Calculate recommendation problems (for activity 5)	
5. Opponent attack recommendation	
Not Intervene-able	The opponent criticizes the player for recommending an intervention on a variable that can't be changed, e.g., "You recommend decreasing people's genetic propensity for weight gain – that's impossible!"
6. Judge: Ask mechanism	
Asks for Mechanism	The judge asks the player to explain how their recommendation affects the outcome, e.g., "How does decreasing junk food advertising decrease obesity?"
7. Student: Pick mechanism	
Pick Mechanism	The student constructs a mechanism using text-based combo boxes, e.g., if the student's recommendation is "decreasing junk food ads", then she might select *decreases* / *junk food eaten*, then selects *decreases / obesity.* The student can add an arbitrary number of additional paths in her explanation.
8. Player: Say mechanism	

Say Mechanism	The player character states the cause the student picked to attack in prose, e.g., "I disagree that junk food ads increase the amount of junk food eaten."

9. Calculate mechanism problems (for activity 10)

10. Opponent: Attack mechanism

Not Intervene-able	See move of same name in activity 5.
Missing Outcome	The opponent criticizes the player's mechanism for not including the outcome, e.g., "You recommended decreasing junk food advertising and explained that junk food advertising decreases junk food eaten, but you didn't even argue that that affects obesity at all!"
Missing Recommendation	The opponent criticizes the player's mechanism for not including the player's recommendation (this is prevented in current text-based interface).
Isolated Variable	The opponent criticizes the player's mechanism for including variables that are not connected to the explanation (this is prevented in the current text-based interface).
Irrelevant Outcome	The opponent criticizes the player's mechanism for explaining the effect of the recommendation on an irrelevant variable, e.g., "You say that junk food advertising only affects the type of junk food eaten, but this has nothing to do with obesity!"
Irrelevant Recommendation	The opponent criticizes the player's mechanism for including an irrelevant recommendation, e.g., "You said that exercise will decrease obesity, but the question is whether or not to decrease junk food advertising!" (note this move only occurs in some problems).
Negligible Causes	The opponent criticizes the player's mechanism for including causes with a negligible effect, e.g., "You said that junk food eaten has only a negligible effect on obesity, so decreasing the amount of junk food eaten won't affect obesity!"
Undesired Outcome	The opponent criticizes the player's mechanism for producing an undesirable outcome, e.g., "You said that we should increase junk food advertising which will increase junk food eaten which will *increase obesity* – that is exactly what we do not want!"

11. Calculate claims to attack (for activity 12)

12. Opponent: Demand evidence

In each pick evidence move, the opponent asks the player to provide evidence for a causal claim in the player's mechanism, e.g., "What evidence do you have that junk food advertising increases the amount of junk food eaten?" The only difference between these moves is that the game may present the opponent's state differently depending on the opponent's confidence.

Attack Non-evidence	In this case, the opponent knows that the student is defending a causal claim for which she has no evidence. The game may present the opponent as extremely confident.
Attack Weak Evidence	In this case, the opponent knows that the student is defending a causal claim for which there is little evidence. The game may present the opponent as very confident.
Attack Effectively Weak Evidence	In this case, the opponent knows that the student is defending a causal claim for which the student hasn't collected much evidence. The game may present the opponent as very confident.
Attack Strong Evidence	In this case, the opponent knows that the student is defending a causal claim for which she has stronger evidence than the opponent. The opponent attacks this cause in the hopes that the student will make a mistake citing her evidence. The game may present the opponent as worried.

13. Student: Pick evidence

Pick Evidence	The student picks reports that support her claim from the list of reports she found earlier in the game.

14. Player: Say evidence

Say Evidence	The player character describes the student's evidence for her causal claim, e.g., "The report *It takes a child to raze a village* shows that violent video games do not increase violent behavior!"
Say No Evidence	The player character describes the student's lack of evidence for her causal claim, e.g., "Uh, … I don't really have any evidence."

15. Calculate evidence quality (for activity 16)

16. Opponent: Judge evidence

Player Cites No Evidence	The opponent criticizes the player for not citing any evidence for the player's causal claim, e.g., [with smug expression] "I see."
Player Cites Irrelevant Report	The opponent criticizes the player for citing evidence that does not support the causal claim, e.g., "*It takes a child to raze a village* does not say anything about the effects of junk food advertising!"
Player Evidence Loses	The opponent cites counter evidence that is superior to the player's evidence, e.g., "While you claim that this newspaper editorial shows that parental permissiveness is the main cause of obesity, these three scientific reports beg to differ!"
Player Evidence Ties	Like *Player Evidence Loses* but in this case the opponent realizes that their evidence is not superior to the students.
Player Evidence Wins	Like *Player Evidence Loses* but in this case student's evidence is superior. The game may present the opponent as worried.

17. Opponent: Propose alternative

Propose alternative	If the player recommends doing nothing, then (rather than ask the player to disprove every possible mechanism) the judge will propose an alternate recommendation and mechanism and ask the player to choose part of that mechanism to disprove, e.g., "Your opponent says that decreasing junk food ads will decrease the amount of junk food eaten which will decrease obesity. Which part of the explanation do you disagree with?"

18. Student: Pick Mechanism attack

Pick Mechanism Attack	If the student has recommended doing nothing and the judge has proposed an alternate recommendation and mechanism, then the student uses a combo box to select a cause in that mechanism to attack, e.g., "junk food ads increase amount of junk food eaten."

19. Opponent: Rebut attack

Alternative cause irrelevant	If the student attacks the opponent's mechanism by criticizing a cause that is not in the opponent mechanism, the opponent will rebut the attack, e.g., "That has nothing to do with my explanation!"

Alternative relation irrelevant	If the student attacks the opponent's mechanism by criticizing a cause that has the same variables as a cause in the opponent's mechanism but not the same causal relation, the opponent will rebut the attack, e.g., "I said exercise *decreases* obesity, not *increases*!"
20. Player: Concede mechanism	
Player Concedes Mechanism	In this case, the player has recommended doing nothing but, when presented with the opponent's alternative recommendation and mechanism, decides that the alternative is correct. This move forfeits the debate.
21. Judge: Finish	
Draw	The judge decides that the player has not beaten their opponent, e.g., "I'm afraid I don't find either of your cases compelling."
Player Loses	The judge decides that the opponent's case is stronger, e.g., "I agree with Mr. Harding's recommendation."
Player Wins	The judge decides that the player's case is stronger, e.g., "Congratulations, you've made an excellent case for banning junk food advertising."

Appendix E. Tutoring rules

Tutoring questions

The pedagogical moves in Policy World are implemented as a stack of question objects. Each question object consists of a prompt, a set of possible student responses, and a method for evaluating the correctness of the student's response. If the student's response is incorrect, the tutor will provide feedback or ask one or more sub-questions. Once the subquestion(s) is answered correctly, the tutor will re-ask the original question. A particular question may contain multiple prompts, only one of which will be asked depending on the specifics of how a particular diagnosis rule has been violated. The question objects are described in the following format:

Prompt: Does smoke cause fire?
Input: <<Yes, No>>
Evaluation: 1. Correct (e.g., student answers "No")
2. Error (e.g., student answers "Yes")
=> QFire (i.e., ask a subquestion about fire)

Note that even simple feedback messages are represented as a question object, where the prompt provides the feedback message, and the only possible student response is to acknowledge the message.

Table E.1

Socratic Tutoring Rules and Moves used in the Policy World Pedagogical Module

Question	
Analysis	
QSearch	If there are un-searched sites, then ask: *You still have [5] more websites to look at! You need to search more!* << OK >> If there are un-searched reports, then ask: *You still have [3] more reports to find! You need to search more!* << OK >> If there are un-searched claims, then ask: *You still have [8] more causal claims to find! You need to analyze your reports!* << OK >>
QQuote	*The text you quoted does not contain a causal claim. Look for causal words like "leads to", "increases", "decreases", "results in".* << OK >>

QEvaluate	If the quote type is incorrect (e.g., student identifies a case study as an experiment), ask: *That quote type is wrong, this is [an experiment].* << OK >> If the quote strength violates evidence ordering, e.g., student rates a case study as being stronger than an experiment, ask: *The strength of that claim is incorrect. [Experiments] provide [stronger] evidence than [case studies].* << OK >> If the quote strength violates evidence consistency, e.g., student rates case studies they disagree with as having a strength of 2, but rates current case that they agree with as 5, ask: *The strength of that claim is incorrect. You rated the other [case studies] as having a strength of 2, not 5.* << OK >>
QDiagramCauseConflictingCitation	*Your citations conflict. Your previous citations indicate that this arrow represents the effect of [junk food advertising] on [the amount of junk food consumed], but your current citation indicates that the arrow represents the effect of [exercise] on [obesity]!* << OK >>
QDiagramCauseNoCitation	*You must link the causal claim you are analyzing to your diagram.* << OK >>
QDiagramCauseRelationWrong	*The relation in your diagram in incorrect. The cause claim indicates that [exercise] [decreases] [obesity] but the arrow you linked to indicates that [exercise] [increases] [obesity]!* << OK >>
QDiagramConflictingVariables	*This quote is about the causal relation between [exercise] and [obesity], but your other citations indicate that the arrow you've linked to from [variable 1] to [variable 2] describes the relation between [advertising] and [junk food consumed].* << OK >>
QDiagramRedundantCause	*This quote is about the effect of [exercise] on [obesity], which is already represented by the arrow from [variable 1] to [variable 2] on your diagram, according to your other citations. You should link this quote to the same arrow, and possibly perhaps rethink your variable names if they are unclear.* <<OK>>
QDiagramUnusedQuote	*You have a causal claim that is no longer linked to your diagram.* << OK >>
QDiagramVariableAmbiguous	*You have an ambiguous variable [fat] in your diagram. Some of your citations indicate that [fat] refers to [obesity], but other citations indicate that [fat] refers to [the amount of junk food consumed].* << OK >>
QDiagramVariableModifierWrong	*[Junk food commercials] is an [intervention]. You should indicate this on your diagram by right-clicking on the variable and choosing "[intervention]".* << OK >>
QDiagramVariableUnconnected	*You have a variable on your diagram ([obesity]) that is not connected to another variable.* << OK >>
QSynthesizeBelief	*Warning! You believe that [exercise] [increases] [obesity], but the majority of evidence indicates that [exercise][decreases][obesity]. You may want to change your belief!* << OK >>
QSynthesizeShift	*Error! You indicated that you are now more certain that [exercise] [increases] [obesity], but the evidence you are currently analyzing says that [exercise] [decreases] [obesity].* << OK >>

QSynthesizeStrength	*Error! You just indicated that [exercise] [increases] [obesity], but your ratings of the evidence about [exercise] and [obesity] suggest that the majority of evidence implies that [exercise] [decreases] [obesity].* << OK >>

Debate

QAffects	*Does [obesity] affect [exercise]? (Where the cause obesity is directly connected to the effect exercise with no mediating variables).* << Yes / No / I don't know >> 1. Correct (yes) 2. Incorrect (no) => QReadArrow
QAffectsOutcome	*Does [the amount of junk food commercials seen] affect [obesity]? (Where the cause is one of the interventions, and the effect is the desired outcome).* << Yes / No / I don't know >> 1. Correct 2. Incorrect => QPath
QChangeAffects	*How does increases/decreasing [exercise] affect [obesity]? (Where the cause obesity is directly connected to the effect exercise with no mediating variables).* << Increase / Decrease >> 1. Correct 2. Incorrect => QReadArrow
QChangeOutcome	*How should you change [the amount of junk food commercials seen] to [decrease] [obesity]?* << Increase / Decrease >> 1. Correct 2. Incorrect => (for each cause, last to first) QChangeAffects
QDescribeEv	*How would you describe this intervention?* << Use the debate interface to pick the list of reports previously identified as a result of QEvidence >> 1. Correct 2. Incorrect
QDescribeMech	*How would you describe this mechanism?* << Use the debate interface to construct a verbal description of the mechanism previously selected from the student's diagram in QMechanism >> 1. Correct 2. Incorrect
QDescribeMechAttack	*How would you describe your attack?* << Use the debate interface to pick which cause in the opponent's mechanism to attack >> 1. Correct 2. Incorrect
QDescribeRec	*How would you describe this intervention?* << Use the debate interface to pick the recommendation previously identified as a result of QRecommendation >> 1. Correct 2. Incorrect
QDesiredChange	*How should the outcome change?* << Increase / Decrease / No change >> 1. Correct 2. Incorrect

QDoesntAffectOutcome	*Does the intervention [amount of junk food commercials seen] affect the outcome [obesity]?* << Yes / No / Don't know >> 1. Correct 2. Incorrect => QNoPath
QEvidence	*Oops, you picked the wrong evidence.* << OK >> 1. OK => QPickEvidenceIndex, QPickEvidence, QDescribeEvidence
QIdInterventionManipulation	*How should we manipulate the intervention?* << Increase / Decrease >> 1. Correct 2. Incorrect => QPath, QDesiredChange, QChangeOutcome
QIdInterventions	*What intervention should we do? Click submit to do nothing.* << Pick diagram objects >> 1. Correct 2. Error: picked non-variable 3. Error: picked non-intervene-able variable 4. Error: picked intervention with no effect => QDoesntAffectOutcome 5. Error: Picks only some (or none) of best interventions => QOtherInterventions, (for each best intervention) QAffectsOutcome
QIdOpponentsCause	*Which arrow on your diagram corresponds to your opponent's claim that [excercise] [increases] [obesity]?* << Pick diagram objects >> 1. Correct 2. Error: picked the wrong arrow 3. Error: picked too many objects 4. Error: picked nothing
QIdOpponentsPath	*The opponent described their mechanism as …. Identify that mechanism on your diagram.* << Pick diagram objects >> 1. Correct 2. Incorrect => QPath
QIdOutcome	*What is the outcome?* << Pick diagram objects >> 1. Correct 2. Error: Pick non-variable 3. Error: Pick non-outcome variable
QIdWeakness	*Which causes on the opponent's path have weak evidence?* << Pick diagram objects >> 1. Correct 2. Incorrect
QMechanism	*Oops, your mechanism is incorrect.* << OK >> 1. OK => QPath, (for each cause) QChangeAffects, QDescribeMech
QMechanismAttack	*Oops, your attack on the opponent's mechanism is incorrect.* << OK >> 1. OK => QIdOpponentsPath, QIdWeakness, QDescribeMechAttack

QNoAffect	*Does the cause affect the effect?* << Yes / No / I don't know>> 1. Correct 2. Error: picked non-belief => QNonBelief 3. Error: picked negligible => QReadArrow
QNonBelief	*Do you believe this relation, according to your diagram?* << Yes / No / Don't know >> 1. Correct (no) 2. Incorrect
QNoPath	*What is the path from the intervention [the amount of junk food commercials seen] to the outcome [obesity]?* << Pick diagram objects >> 1. Correct - no path 2. Missing Intervention 3. Missing outcome 4. Unconnected path For each negligible/non belief cause => QNoAffects
QOtherInterventions	*What other interventions might you do?* <<Pick diagram objects >> 1. Correct 2. Error: Pick non-variable 3. Error: Pick non-intervention 4. Error: Misses interventions
QPath	*What is the path from [the number of junk food commercials seen] to [obesity]?* <<Pick diagram objects>> 1. Correct 2. Missing intervention 3. Missing outcome 4. Error: No path or unconnected path => (for each cause in correct path) QAffects 5. Error: Bad path => (for each bad cause) QNoAffects
QPickEvidence	*Which citations support your claim that [exercise] [increases] [obesity]?* << Pick citations from list of citations linked to given cause >> 1. Correct 2. Incorrect
QPickEvidenceIndex	*To find the citations that support your claim that [exercise] [increases] [obesity], click on that arrow.* << Pick diagram objects >> 1. Correct 2. Incorrect
QReadArrow	*What symbol do you seen on the highlighted arrow?* << + / 0 / - >> 1. Correct 2. Incorrect
QRecommend	*Oops, your recommendation is incorrect.* << OK >> 1. OK => QIdOutcome, QIdInterventions, (if there is a best intervention) QIdInterventionManipulation, QDescribe

Appendix F: Intrinsic Motivation Inventory

The Intrinsic Motivation Inventory (IMI) "is a multidimensional measurement device intended to assess participants' subjective experience related to a target activity in laboratory experiments" (University of Rochester, 2008). The questions below adapted from IMI were used to measure interest in Chapter 6. Each question is rated on a 7-point Likert scale where an (R) indicates the question is reverse scored.

Interest/Enjoyment

1. I enjoyed playing this game very much.
2. This game was fun to play.
3. I thought this was a boring game. (R)
4. This game did not hold my attention at all. (R)
5. I would describe this game as very interesting.
6. I thought this game was quite enjoyable.
7. While I was playing this game, I was thinking about how much I enjoyed it.

Perceived competence

8. I think I am pretty good at this game.
9. I think I did pretty well at this game, compared to other students.
10. After playing this game for awhile, I felt pretty competent.
11. I am satisfied with my performance in this game.
12. I was pretty skilled at this game.
13. This was a game that I couldn't do very well. (R)

Effort / Importance

14. I put a lot of effort into this.
15. I didn't try very hard to do well at this game. (R)
16. I tried very hard on this game.
17. It was important to me to do well at this game.
18. I didn't put much energy into this. (R)

Pressure/tension

19. I did not feel nervous at all while playing this. (R)
20. I felt very tense while playing this game.
21. I was very relaxed in playing this game. (R)
22. I was anxious while playing this game.

23. I felt pressured while playing this game.

Perceived choice

24. I believe I had some choice about taking different actions in the game.
25. I felt like it was not my own choice to take different actions in the game. (R)
26. I didn't really have a choice about taking different actions in the game. (R)
27. I felt like I had to do different actions in the game.
28. I took different actions in this game, because I had no choice. (R)
29. I took different actions because I wanted to. (R)
30. I took different actions in this game because I had to. (R)

Value / usefulness

31. I believe this activity could be of some value to me.
32. I think that doing this activity is useful for learning about policy.
33. I think this is important to do, because it can teach you about policy
34. I would be willing to do this again, because it has some value to me.
35. I think doing this activity could help me to learn about policy
36. I believe doing this activity could be beneficial to me.
37. I think this is an important activity.

Bibliography

Ainsworth, S. E. (2006). Deft: A conceptual framework for considering learning with multiple representations. *Learning and Instruction*, *16*(3), 183-198.

Axelrod, R. (1976). *Structure of decision: The cognitive maps of political elites*. Princeton University Press Princeton, NJ.

Ball, D. L., & Forzani, F. M. (2007). What makes educational research "educational"? *Educational Researcher*, 36(9), 529-540.

Bauer, M. I., & Johnson-Laird, P. N. (1993). How diagrams can improve reasoning. *Psychological Science*, *4*(6), 372-378.

Beardsley, M. C. (1950). *Practical logic.* New York: Prentice Hall.

Bernstein (1992). *Euclid: Supporting collaborative argumentation with hypertext* (Tech. Rep. No. CU-CS-596-92). University of Colorado at Boulder.

Bredeweg, B., Bouwer, A., Jellema, J., Bertels, D., Linnebank, F. F., & Liem, J. (2006). Garp3: A new workbench for qualitative reasoning and modelling. In C. Bailey-Kellogg, & B. Kuipers (Eds.), *Proceedings of the 20th International Workshop on Qualitative Reasoning.* (pp. 21-8). Hanover, NH.

Brem, S. K., Russell, J., & Weems, L. (2001). Science on the web: Student evaluations of scientific arguments. *Discourse Processes*, *32*(2), 191-213.

Britt, M. A., & Aglinskas, C. (2002). Improving students' ability to identify and use source information. *Cognition & 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.

Buckingham Shum, S. J., Selvin, A. M., Sierhuis, M., Conklin, J., Haley, C. B., & Nuseibeh, B. (2006). Hypermedia support for argumentation-based rationale: 15 years on from gIBIS and QOC. In A. H. Dutoit, R. McCall, I. Mistrik, & B. Paech (Eds.), *Rationale management in software engineering.* (pp. 111-32). Berlin: Springer-Verlag.

Buckingham Shum, S., Uren, V., Li, G., Sereno, B., & Mancini, C. (2007). Modeling naturalistic argumentation in research literatures: Representation and interaction design issues. *International Journal of Intelligent Systems*, *22*(1), 17-47.

Bush, V. (1945). *Science, the endless frontier.* Washington D.C.: United States Government Printing Office.

Cañas, A. J., Hill, G., Carff, R., Suri, N., Lott, J., Eskridge, T., et al. (2004). CmapTools: A knowledge modeling and sharing environment. In A. J. Cañas, J. D. Novak, & F. M. González (Eds.), *Concept maps: Theory, methodology, technology. Proceedings of the First International Conference on Concept Mapping*. (pp. 125-33). Pamplona, Spain: Universidad Pública de Navarra.

Capcom (2005). Phoenix wright: Ace attorney. [Computer Software] Osaka, Japan: Capcom.

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.

Causal Mapping for Research in Information Technology. (2005). *Causal mapping for research in information technology.* Hershey, PA: Idea Group.

Chinn, C. A., & Brewer, W. F. (1998). An empirical test of a taxonomy of responses to anomalous data in science. *Journal of Research in Science Teaching*, *35*(6), 623-654.

Chinn, C. A., & Brewer, W. F. (2001). Models of data: A theory of how people evaluate data. *Cognition and Instruction*, *19*(3), 323-393.

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education*, *86*(2), 175-218.

Chryssafidou, E. (2000). DIALECTIC: Enhancing essay writing skills with computer-supported formulation of argumentation. In C. Stephanidis (Ed.), *Proceedings of the ERCIM WG UI4ALL one-day joint workshop with i3 Spring Days 2000 on "Interactive learning environments for children"*. (14 pages). Athens, Greece.

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

Collins, A. (1976). *Processes in acquiring knowledge* (Tech. Rep. No. 3231). Cambridge, MA: Bolt Beranek and Newman Inc.

Collins, A. (1977). Processes in acquiring knowledge. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the acquisition of knowledge.* (pp. 339-63). Hillsdale, NJ: Lawrence Erlbaum.

Computing Research Association (2005). *Cyberinfrastructure for education and learning for the future: A vision and research agenda*. Retrieved from http://www.cra.org/uploads/documents/resources/rissues/cyberinfrastructure.pdf

Computing Research Association (2003). *Grand research challenges in information systems*. Retrieved from http://www.cra.org/uploads/documents/resources/rissues/gc.systems_.pdf

Corbett, A. T., & Anderson, J. R. (2001). Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. In J. Jacko, A. Sears, M. Beaudouin-Lafon, & R. Jacob (Eds.), *Proceedings of the ACM CHI '2001 Conference on Human Factors in Computing Systems*. (pp. 245-52). New York: ACM Press.

Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, *88*(4), 715-730.

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

Cox, R. (1997). Representation interpretation versus representation construction: An ILE-based study using switchERII. In B. du Boulay, & R. Mizoguchi (Eds.), *Proceedings of the 8th World Conference on Artificial Intelligence in Education*. (pp. 434-41). Amsterdam: IOS.

Druzdzel, M. J. (1999). SMILE: Structural modeling, inference, and learning engine and genie: A development environment for graphical decision-theoretic models. In *Proceedings of the 16th National Conference on Artificial Intelligence*. (pp. 902-3). Menlo Park, CA: American Association for Artificial Intelligence.

Easterday, M. W. (in press). Policy world: A cognitive game for teaching deliberation. In N. Pinkward, & B. McLaren (Eds.), *Educational technologies for teaching argumentation skills.* Oak Park, IL: Bentham Science Publishers.

Easterday, M. W., Aleven, V., & Scheines, R. (2007a). 'Tis better to construct than to receive? The effects of diagram tools on causal reasoning. 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. (2007). 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). Marina Del Rey, CA.

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*. (pp. 317-23). Stanford, CA: CSLI Publications.

Easterday, M. W., Aleven, V., Scheines, R., & Carver, S. M. (in press). Constructing causal diagrams to learn deliberation. *International Journal of Artificial Intelligence in Education.*

Edelson, D. C., & Reiser, B. J. (2006). Making authentic practices accessible to learners: Design challenges and strategies. In R. K. Sawyer (Ed.), *Cambridge handbook of the learning sciences.* (pp. 335-54). New York: Cambridge University Press.

Filament Games. (in devlopment). *Guardian of law*. [Computer Software]. Madison, WI: Filament Games.

Filament Games. (2010). *Argument Wars* [Computer Software]. Washington, DC: iCivics.

Fischer, G., McCall, R., & Morch, A. (1989). JANUS: Integrating hypertext with a knowledge-based design environment. In R. Akscyn (Ed.), *Proceedings of the 2nd Annual ACM Conference on Hypertext*. (pp. 105 - 117). New York: ACM.

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.

Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. *Simulation & Gaming*, *33*(4), 441.

Gobert, J. D. (2005). The effects of different learning tasks on model-building in plate tectonics: Diagramming versus explaining. *Journal of Geosicence Education*, *53*(4), 444-455.

Gordon, & Walton (2003). The Carneades argumentation framework: Using presumptions and exceptions to model critical questions. In P. E. Dunne, & T. J. M. Bench-Capon (Eds.), *Proceedings of the 2006 Conference on Computational Models of Argument.* (pp. 195-207). Amsterdam, The Netherlands: IOS Press.

Gore, A. (2007). *The assault on reason.* New York: Penguin Press.

Graesser, A. C., & Bertus, E. L. (1998). The construction of causal inferences while reading expository texts on science and technology. *Scientific Studies of Reading*, *2*(3), 247-269.

Graesser, A. C., & Olde, B. A. (2003). How does one know whether a person understands a device? The quality of the questions the person asks when the device breaks down. *Journal of Educational Psychology*, *95*, 524-536.

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal*, *31*(1), 104-137.
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 and 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), 168-182.
Hair, D. C. (1991). Legalese: A legal argumentation tool. *SIGCHI Bulletin*, 23(1), 71-74.
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.
Harrell, M. (2005). Using argument diagramming software in the classroom. *Teaching Philosophy*, *28*(2), 163-77.
Hastings, P., Britt, A., Sagarin, B., Durik, A., & Kopp, K. (2009). Designing a game for teaching argumentation skills. In C. H. Lane, A. Ogan, & V. Shute (Eds.), *Proceedings of the Workshop on Intelligent Educational Games at the 14th International Conference on Artificial Intelligence in Education.* (pp. 21-30). Brighton, UK.
Heuer, R. J. (1999). *Psychology of intelligence analysis*. New York: Novinka Books.
Hogwood, B. W., & Gunn, L. A. (1984). *Policy analysis for the real world.* Oxford University Press.
Horn, R. E., & Weber, R. P. (2007). *New tools for resolving wicked problems: Mess mapping and resolution mapping processes*. Strategy Kinetics L.L.C. Retrieved from http://www.strategykinetics.com/files/New_Tools_For_Resolving_Wicked_Problems.pdf
International Artificial Intelligence in Education Society (n.d.). *Journal scope and standards*. Retrieved from http://ijaied.org/journal/scope/
Irwin, L. G. (2003). *The policy analyst's handbook: Rational problem solving in a political world*. Armonk, NY: M.E. Sharpe.
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.
Karacapilidis, & Papadias (2001). Computer supported argumentation and collaborative decision making: The HERMES system. *Information Systems*, 26(4), 259-277.
Kelly, K. K. (2010, July 28). Master planner: Fred brooks shows how to design anything. *Wired magazine*. Retrieved from: http://www.wired.com/magazine/2010/07/ff_fred_brooks/
Kim, S., Taber, C. S., & Lodge, M. (2009). A computational model of the citizen as motivated reasoner: Modeling the dynamics of the 2000 presidential election. *Political Behavior*, 32(1), 1-28.
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.
Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75-86.
Klahr, D. (2000). *Exploring science: The cognition and development of discovery processes*. Cambridge, MA: MIT Press.
Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors . *Educational Psychology Review*, *19*(3), 239-264.
Koedinger, K. R., & Corbett, A. (2006). Cognitive tutors: Technology bringing learning science to the classroom. In K. Sawyer (Ed.), *The cambridge handbook of the learning sciences.* (pp. 61-78). *Cambridge University Press*.
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), 129-164.
Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, *8*(1), 30-43.
Kornbluh, M., Fegan, M., & Rehberger, D. (2005). Media matrix: A digital library research tool. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries.* Denver, Colorado, USA.
Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. Cambridge, MA: MIT Press.
Kozma, R. B., & Russell, J. (1997). Multimedia and understanding: Expert and novice responses to different representations of chemical phenomena. *Journal of Research in Science Teaching*, *34*(9).
Kuhn, D. (1991). *The skills of argument.* New York: Cambridge University Press.

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), 261-288.
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), 363-385.
Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480-498.
Kuypers, J. A. (2009). Framing analysis. In J. A. Kuypers (Ed.), *Rhetorical criticism: Perspectives in action.* (pp. 181-204). Lexington Books.
Lakoff, G. (2002). Moral politics: How liberals and conservatives think. Chicago: University of Chicago Press.
Lane, C. H., Hays, M. J., Auerbach, D., & Core, M. G. (2010). Investigating the relationship between presence and learning in a serious game. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Proceedings of the 10th International Conference on Intelligent Tutoring Systems.* (pp. 274-84). Berlin: Springer.
Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, *11*, 65-99.
Lau, R. R., & Redlawsk, D. P. (2001). Advantages and disadvantages of cognitive heuristics in political decision making. *American Journal of Political Science*, *45*(4), 951-971.
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.
Lodge, M., McGraw, K., & Stroh, P. (1989). An impression-driven model of candidate evaluation. *American Political Science Review*, *83*(2), 399-419.
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), 2098-2109.
Lynch, C. F., Ashley, K. D., Aleven, V. A., & Pinkwart (2006). Defining "ill-defined domains": A literature survey. In V. A. Aleven, K. D. Ashley, C. F. 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*. Johngli, Taiwan: National Central University.
MacCoun, R. J. (1998). Biases in the interpretation and use of research results. *Annual Reviews in Psychology*, *49*(1), 259-287.
Maki, R. H. (1998). Test predictions over text material. In D. J. Hacker, A. C. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice.* (pp. 117-44). Mahwah, NJ: Lawrence Erlbaum.
Mapping Strategic Knowledge. (2002). *Mapping strategic knowledge*. London: Sage.
Marshall, C. C. (1987). Exploring representation problems using hypertext. In *Proceedings of the ACM Conference on Hypertext.* New York, NY: ACM Press.
Marshall, Halasz, Rogers, & Janssen Jr. (1991). Aquanet: A hypertext tool to hold your knowledge in place. In *Proceedings of the Third Annual ACM Conference on Hypertext*. (pp. 261-75). San Antonia, TX.
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.
Mathan, S. A., & Koedinger, K. R. (2005). Fostering the intelligent novice: Learning from errors with metacognitive tutoring. *Educational Psychologist*, *40*(4), 257-265.
Mayer, & Moreno (2002). Aids to computer-based multimedia learning. *Learning and Instruction*, *12*(1), 107-119.
Mayer, R. E. (2001). *Multimedia learning.* UK: 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), 367-388.
McElheny, V. K. (2002). Biographical memoirs: Edwin Herbert Land. *Proceedings of the American Philosophical Society*, *146*(1), 111-122.
McManus, J. (2009). *Detecting bull: How to identify bias and junk journalism in print, broadcast and on the wild web.* Sunnydale, CA: The Unvarnished Press.
Metcalf, S. J., Krajcik, J., & Soloway, E. (2000). Model-It: A design retrospective. In M. J. Jacobson, & R. B. Kozma (Eds.), *Innovations in science and mathematics education: Advanced design for technologies of learning.* (pp. 77-115). Mahwah, NJ: Lawrence Erlbaum.
Mitrovic, A., Mayo, M., Suraweera, P., & Martin, B. (2001). Constraint-Based tutors: A success story. In L. Monostori, J. Váneza, & M. Ali (Eds.), *Lecture notes in computer science: Vol. 2070. Engineering of intelligent systems.* (pp. 931-40). Berlin: Springer.

Molden, D. C., & Higgins, E. T. (2005). Motivated thinking. In K. J. Holyoak, & R. G. Morrison (Eds.), *The cambridge handbook of thinking and reasoning.* (pp. 295-317). New York: Cambridge University Press.
Montibeller, G., & Belton, V. (2006). Causal maps and the evaluation of decision options-a review. *Journal of the Operational Research Society*, *57*(7), 779-791.
Montibeller, G., Belton, V., Ackermann, F., & Ensslin, L. (2008). Reasoning maps for decision aid: An integrated approach for problem-structuring and multi-criteria evaluation. *Journal of the Operational Research Society*, *59*(5), 575.
Mott, B. W., & Lester, J. C. (2006). Narrative-Centered tutorial planning for inquiry-based learning environments. In M. Ikeda, K. Ashley, & T. -W. Chan (Eds.), *Lecture notes in computer science: Vol. 4053. Intelligent tutoring systems.* (pp. 675-84). Berlin: Springer.
Nathan, M. J., & Alibali, M. W. (2010). Learning sciences. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(3), 329-345.
Nathan, M. J. (1998). Knowledge and situational feedback in a learning environment for algebra story problem solving. *Interactive Learning Environments*, *5*(1), 135-159.
National Research Council (1999). *How people learn: Bridging research and practice*. Washington, DC: National Academy Press.
Nelson, B. C., Ketelhut, D. J., & Schifter, C. (2009). Embedded assessments of science learning in immersive educational games: The SAVE science project. In C. H. Lane, A. Ogan, & V. Shute (Eds.), *Proceedings of the Workshop on Intelligent Educational Games at the 14th International Conference on Artificial Intelligence in Education.* (pp. 121-4). Brighton, UK.
Nesbit, C., & Adesope, O. (2006). Learning with concept and knowledge maps: A meta-analysis. *Review of Educational Research*, *76*(3), 413-48.
Newman, S. E., & Marshall, C. C. (1992). *Pushing toulmin too far: Learning from an argument representation scheme* (Tech. Rep. No. SSL-92-45). Palo Alto, CA: Xerox PARC.
Newman, W. (1994). A preliminary analysis of the products of HCI research, using pro forma abstracts. In B. Adelson, S. Dumais, & J. Olson (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* (pp. 278-84). New York: ACM.
Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175-220.
Norton, D. (2009, May 7). *Filament games development blog*. Retrieved from http://www.filamentgames.com/node/250
Novick, L. R., & Hurley, S. M. (2001). To matrix, network, or hierarchy: That is the question. *Cognitive Psychology*, *42*(2), 158-216.
Patel, V. L., & Groen, G. J. (1986). Knowledge based solution strategies in medical reasoning. *Cognitive Science*, *10*(1), 91-116.
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). Mahwah, NJ: Lawrence Erlbaum.
Pinkwart, N., Aleven, V., Ashley, K., & Lynch, C. (2006). Toward legal argument instruction with graph grammars and collaborative filtering techniques. In M. Ikeda, K. Ashley, & T. W. Chan (Eds.), *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*. Berlin, Germany: Springer.
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.
Pirolli, P., & Card, S. (2005). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of the International Conference on Intelligence Analysis*. Retrieved from https://analysis.mitre.org/proceedings/Final_Papers_Files/206_Camera_Ready_Paper.pdf
Redlawsk, D. P. (2002). Hot cognition or cool consideration? Testing the effects of motivated reasoning on political decision making. *The Journal of Politics*, *64*(4), 1021-1044.
Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences*, *4*, 155-169.
Rolf, B., & Magnusson, C. (2002). Developing the art of argumentation. A software approach. In *Proceedings of the 5th International Conference of the International Society for the Study of Argument*. Amsterdam: SIC SAT.
Rowe, G., & Reed, C. (2006). Translating Wigmore diagrams. In *Proceedings of the 2006 Conference on Computational Models of Argument*. (pp. 171-82). Amsterdam, The Netherlands: IOS Press.

Salles, P., Bredeweg, B., & Araújo, S. (2006). Qualitative models about stream ecosystem recovery: Exploratory studies. *Ecological Modelling*, *194*(1-3), 80-89.
Sandell, M. (2009). *Justice with Michael Sandel*. WGBH Educational Foundation and the President and Fellows of Harvard College. Retrieved from http://www.justiceharvard.org/
Sandoval, W. A. (2003). Conceptual and epistemic aspects of students' scientific explanations. *Journal of the Learning Sciences*, 12(1), 5-51.
Scardamalia, M., & Bereiter, C. (2006). Knowledge building: Theory, pedagogy, and technology. *The Cambridge Handbook of the Learning Sciences*, 97-115.
Schank, & Ranney (1995). Improved reasoning with convince me. In I. Katz, R. Mack, & L. Marks (Eds.), *Conference Companion on Human Factors in Computing Systems*. (pp. 276-7). New York: ACM.
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.
Scheuer, O., Loll, F., Pinkwart, N., & McLaren, B. M. (2010). Computer-Supported argumentation: A review of the state of the art. *International Journal of Computer-Supported Collaborative Learning*, *5*(1).
Scheuer, O., McLaren, B. M., Loll, F., & Pinkwart, N. (n.d.). Automated analysis and feedback techniques to support argumentation: A survey. In N. Pinkwart, & B. McLaren (Eds.), *Educational technologies for teaching argumentation skills.* Oak Park, IL: Bentham Science Publishers.
Schumacher, R. M., & Gentner, D. (1988). Transfer of training as analogical mapping. *IEEE Transactions on Systems, Man and Cybernetics*, *18*(4), 592-600.
Schunk, D. H., Pintrich, P. R., & Meece, J. L. (2008). *Motivation in education: Theory, research, and applications* (3rd ed.). Upper Saddle River, NJ: Pearson.
Schwenk, C. R. (1995). Strategic decision making. *Journal of Management*, *21*(3), 471.
Serious Games Interactive (2008). *Global conflicts: Latin america* [Computer Software]. Retrieved from http://www.seriousgames.dk/
Sherwin, C. W., & Isenson, R. S. (1967). Project hindsight: A defense department study of the utility of research. *Science*, *156*, 1571-1577.
Simon, H. A. (1973). The structure of ill structured problems. *Artificial Intelligence*, *4*(3), 181-201.
Simon, H. A. (1996a). *The sciences of the artificial* (3rd ed.). Cambridge, MA: MIT Press.
Simon, H. A. (1996b). Social planning: Designing the evolving artifact. In H. A. Simon (Ed.), *The sciences of the artificial (3rd ed.).* Cambridge, MA: MIT Press.
Smolensky, P., Fox, B., King, R., & Lewis, C. (1988). Computer-Aided reasoned discourse or, how to argue with a computer. In R. Guindon (Ed.), *Cognitive science and its applications for human-computer interaction*. (pp. 109-62). Hillsdale, NJ: Lawrence Erlbaum.
Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search (2nd ed.).* Cambridge, MA: MIT Press.
Stevens, & Collins (1977). The goal structure of a socratic tutor. In *Proceedings of the 1977 Annual Conference*. New York: ACM.
Stokes, D. (1997). *Pasteur's quadrant: Basic science and technological innovation*. Washington, DC: Brookings Institution Press.
Stull, & Mayer (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), 808-820.
Styvers, M., Tenenbaum, J. B., Wagenmakers, E., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*, 453-489.
Suthers, D. D. (2003). Representational guidance for collaborative inquiry. In J. Andriessen, M. Baker, & D. D. Suthers (Eds.), *Arguing to learn: Confronting cognitions in computer-supported collaborative learning environments.* (pp. 27-46). Dordrecht, The Netherlands: Kluwer Academic.
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. 266-73). Charlottesville, VA: American 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), 755-769.
Tetrad. (2008). *Tetrad* [Computer Software]. Retrieved from http://www.phil.cmu.edu/projects/tetrad/
The JASON Project (2007). *Operation: Resilient planet* [Computer Software] Retrieved from http://www.jason.org

Thille, C. (2008). Creating open learning as a community based research activity. In T. Iiyoshi, & V. Kumar (Eds.), *Opening up education: The collective advancement of education through open technology, open content, and open knowledge*. (pp. 165-79). Cambridge, MA: MIT Press.

Thompson, C. (2007). Halo 3: How Microsoft Labs invented a new science of play. Wired Magazine, 15(09). Retrieved from http://www.wired.com/gaming/virtualworlds/magazine/15-09/ff_halo

Trabasso, T. (2005). The role of causal reasoning in understanding narratives. In T. Trabasso, J. Sabatini, D. W. Massaro, & R. C. Calfee (Eds.), *From orthography to pedagogy: Essays in honor of Richard L. Venezky*. (pp. 81-107). Mahwah, NJ: Lawrence Erlbaum.

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

University of Rochester (2008). *Intrinsic motivation inventory*. Retrieved from http://www.psych.rochester.edu/SDT/measures/IMI_description.php

van den Braak, S. W., van Oostendorp, H., Prakken, H., & Vreeswijk, G. A. W. (2006). A critical review of argument visualization tools: Do users become better reasoners? In *Proceedings of the ECAI-06 Workshop on Computational Models of Natural Argument.* (pp. 67-75). Trento, Italy.

van den Braak, S. W., Vreeswijk, G. A. W., & Prakken, H. (2007). Avers: An argument visualization tool for representing stories about evidence. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*. (pp. 11-5). New York: ACM.

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.

Van Meter, P. (2001). Drawing construction as a strategy for learning from text. *Journal of Educational Psychology*, *93*(1), 129-140.

VanLehn, K. (1985). *Theory reform caused by an argumentation tool* (Tech. Rep. No. ISL-1 1). Palo Alto, CA: Xerox PARC.

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, 16(3), 227-265.

Verheij, B. (1998). Argue! An implemented system for computer-mediated defeasible argumentation. In H. La Poutré, & J. van den Herik (Eds.), *Proceedings of the 10th Netherlands/Belgium Conference on Artificial Intelligence*. (pp. 57-66). Amsterdam, The Netherlands: CWI.

Verheij, B. (1998). Argumed - A template-based argument mediation system for lawyers. In J. C. Hage, T. J. M. Bench-Capon, A. W. Koers, C. N. J. de Vey Mestdagh, & C. A. F. M. Grütters (Eds.), *Proceedings of the 11th Conference on Legal Knowledge Based Systems*. (pp. 113-30). Nijmegen: Gerard Noodt Instituut.

Voss, J. F. (2005). Toulmin's model and the solving of ill-structured problems. *Argumentation*, *19*(3), 321-329.

Voss, J. F., & Wiley, J. (2006). Expertise in history. In K. A. Ericsson, N. Charness, P. Feltovich, & R. R. Hoffman (Eds.), *The cambridge handbook of expertise and expert performance.* (pp. 1746-2424). Cambridge, UK: Cambridge University Press.

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, NJ: Lawrence Erlbaum.

Voss, J. F., Greene, T. R., Post, T. A., & Penner, B. C. (1983). Problem solving skill in the social sciences. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory.* (pp. 165-213). New York: Academic Press.

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. *Public policy analysis: A brief definition* (Document. No. P-7856). Santa Monica, CA: RAND Corporation.

Walton, D. (2006). *Fundamentals of critical argumentation*. New York: Cambridge University Press.

Westen, D. (2007). *The political brain: The role of emotion in deciding the fate of the nation*. New York: Public Affairs.

White, B. Y. (1993). Thinkertools: Causal models, conceptual change, and science education. *Cognition and Instruction*, *10*(1), 1-100.

Winn, B. (1987). Charts, graphs, and diagrams in educational materials. In D. M. Willows, & H. A. Houghton (Eds.), *The psychology of illustration: Volume 1: Basic research*. (pp. 152-98). New York: Springer.

Woolf, B. P. (2009). *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Burlington, MA: Morgan Kaufmann

Zimmerman (2000). The development of scientific reasoning skills. *Developmental Review*, *20*(1), 99-149.

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

Glossary

Analysis tools, part of the inquiry environment, refer to representation tools that partially automate the process of constructing an external representation, e.g., a tool that allows the student to select a variable to produce a bar chart.

Assistance, part of the tutoring platform, is the goal of tutoring, i.e., an intelligent tutor provides feedback, hints and other services to help the student learn how to solve problems.

Authoring tools, part of the tutoring platform, allow instructors to create content.

Causal beliefs, part of the deliberation framework, define the student's mental model of the policy problem, for example that "exercise decreases obesity."

Challenge, part of the motivation model, describes how easy or difficult an activity is for the student. Activities with an *optimal* level of challenge (neither too easy nor too difficult) are thought to be more motivating. See Garris, Ahlers, & Driskell, (2002).

Cognition, in terms of the motivation model, describes the mental processes used during problem solving, e.g., attention.

Cognitive model, part of the tutoring platform, defines the knowledge, skills and dispositions that allow expert performance and which students are supposed to acquire.

Comprehend, part of the deliberation framework, refers to the process of identifying the type and value of different pieces of information, for example in the text: "exercise decreases obesity," *exercise* is the *cause, obesity* is the *effect*, and *decreases* is the *causal relation.*

Construct, part of the deliberation framework, refers to the process of creating an external representation.

Control, part of the motivation model, refers to the amount of choice (or perceived choice) that students have to direct the activity. For example, watching a video provides little control, discovery learning provides a large amount of control, and a video game provides some intermediate level of control, which may be illusory, e.g., they game might allow the player to choose one of two doors, both of which lead to the same outcome. More control is though to increase motivation, and possibly increase floundering. See Garris, Ahlers, & Driskell, (2002).

Curriculum, part of the tutoring platform, refers to the entire body of instruction provided on a given subject.

Decide (via interpretation), part of the deliberation framework, refers to the process of making an inference from an external representation in order to produce a policy recommendation.

Decide (via memory), part of the deliberation framework, refers to the process of making an inference based only on one's memory in order to produce a policy recommendation.

Delayed feedback, part of the tutoring behavior framework, refers to the withholding of feedback until the end of a problem. See VanLehn (2006).

Deliberation, part of the deliberation framework, refers to the process of solving a policy problem and includes questioning, search, comprehension, evaluation, synthesis, construction, and decision.

Delivery system, part of the tutoring platform, refers to software used to distribute and provide access to instruction, for example, the Open Learning Initiative is a delivery system (which also includes authoring tools).

Demand feedback, part of the tutoring behavior framework, refers to delaying feedback or instruction until the student asks for help. See VanLehn (2006).

Diagram, part of the deliberation framework, refers to an external representation of the problem.

Error-general feedback, part of the tutoring behavior framework, refers to feedback on a student's action that indicates why the action was incorrect (but containing no information about the student's misconception that may have caused the error).

Error-specific feedback, part of the tutoring behavior framework, refers to feedback on a student's action that indicates something about the student's misconception that caused the error. See VanLehn (2006).

Evaluate, part of the deliberation framework, refers to how strong the student thinks a piece of evidence is, for example, students should usually evaluate a randomized controlled trial as a stronger piece of evidence than a claim by Aunt Louise (all things being equal).

Fantasy, part of the motivation model, refers to an imaginary, or simulated problem context, e.g., in Carmen Sandiego, the student plays the role of a detective. Fantasy is thought to increase motivation. See Garris, Ahlers, & Driskell, (2002).

Focus, part of the deliberation framework, refers to the particular question, or sub-question that the student is trying to answer, e.g., "What causes childhood obesity?" or "What evidence would prove that exercise increases obesity?"

Goals, part of the motivation model, refers to the degree to which the goal is clear to the student and to which they can judge their progress or distance to the goal. Clear goals and feedback are thought to increase both motivation and learning. See Garris, Ahlers, & Driskell, (2002).

Immediate feedback, part of the tutoring behavior framework, refers to feedback that is provided as soon as the student makes an error. See VanLehn (2006).

Indirect feedback, part of the tutoring behavior framework, refers to changes in the problem situation that may allow the student to make inferernces about the correctness of their actions, or the reasons for an error. For example, in the courtroom game Phoenix Wright, the student may infer that they have made a correct argument when the witness testimony advances – the game does not explicitly tell the student that their argument is correct, and in fact, the student could be heading down a dead end.

Inner loop, part of the tutoring behavior framework, a intelligent tutor responds to each step of the student's problem solving for example by evaluating the step, providing feedback and hints, see VanLehn (2006).

Instructional dynamics, the interactions between students, teachers, and content that define teaching and learning, and which are the proper study of educational research, see Ball and Forzani (2007).

Intelligent novice feedback, part of the tutoring behavior framework, refers to the delaying of feedback until after the student has had an opportunity to observe situational changes resulting from their actions; feedback is only provided when the student fails to perceive an error from the indirect, situational feedback and tries to proceed to the next step. Intelligent novice feedback is thought to sometimes be better than immediate feedback. See Mathan (2005).

Interest, part of the motivation model, refers to the student's relatively stable liking of a particular domain (as opposed to the short-term intrinsic pleasure of a particular task). Interest is thought to increase motivation. See Schunk, Pintrich & Meece (2008).

Intrinsic pleasure, part of the motivation model, refers to the student's immediate liking of a particular task. Intrinsic pleasure is thought to increase motivation, and be subject to manipulation. See Schunk, Pintrich & Meece (2008).

Knowledge-based feedback / hint / explanation, part of the tutoring behavior framework, refers to an explicit explanation of cognitive strategy.

Learning challenges, part of the tutoring platform, refers to difficulties students have in learning the cognitive model.

Microworld, part of the inquiry environment, refers to simulations, data sets, games, or sensors that allow the students to take actions and/or make observations. Microworlds allow students to practice search.

Minimal feedback / Error flagging, part of the tutoring behavior framework, refers to feedback that indicates that an action was correct/incorrect, but does not provide any other information. See VanLehn (2006).

Modeling tools, part of the inquiry environment, are representation tools that allow the student to create computational simulations, e.g., Stella, VModel, and GARP allow the student to create causal diagrams that can then be turned into computational models that then simulate how different variables will change. Modeling tools partially automate the process of diagram interpretation, because they help the student to make inferences about the representation.

Motivation refers to the student's choice of activity, their effort, and their persistence. Motivation is not directly observable but inferred from these behavioral measures. See Schunk, Pintrich & Meece (2008).

Multiple representations refers to the problem in ill-defined domains that different problem-solvers may choose different representational systems for describing a problem, or create different particular representations of the problem within the same representational system. Finding a representational system that improves performance in a given ill-defined domain reduces this problem.

Mystery, part of the motivation model, refers to the information complexity of a game, i.e., a perceived inconsistency or gap in one's knowledge. An optimal level of mystery is thought to increase intrinsic motivation by evoking curiosity. See Garris, Ahlers, & Driskell, (2002).

Problem, a.k.a. task, refers to the question the student must solve. Intelligent tutors typically teach problem-solving.

Inquiry environment, part of the tutoring platform, refers to the components that allow problem solving to take place, e.g., microworlds, recording tools, representation tools, and process scaffolding. Intelligent tutors provide assistance within some inquiry environment.

Processed information, part of the deliberation framework, refers to the schematized output of comprehension.

Questioning, part of the deliberation framework, refers to the process of setting the current focus of problem solving.

Raw information, part of the deliberation framework, refers to the output of search, e.g. observations, reports, background knowledge, etc.

Recommendation, part of the deliberation framework, refers to the student's solution to the policy problem, e.g., if the problem is: *Should we decrease junk food advertising?*, the student's recommendation might be "Yes." Recommendations may of course be more complex, and also require explanation of why the recommendation will achieve a desired outcome, or an argument about why one recommendation is better than another.

Recording tools, part of the inquiry environment, refer to tools that allow the student to collect, annotate, or store information.

Representation tools, part of the inquiry environment, allow the student to create external representations of the problem, e.g., an argument diagramming tool. Analysis and modeling tools are types of representation tools.

Search, part of the deliberation framework, refers to the process of finding information in order to solve the problem. Search includes everything from consulting background knowledge, using Google to find reports, or conducting an experiment.

Self-efficacy, part of the motivation model, refers to the student's beliefs about how competent they are at a given activity. Self-efficacy is thought to be part of a virtuous motivational cycle where learning increases self-efficacy, which increases intrinsic pleasure, which increases motivation, which increases learning. See Schunk, Pintrich & Meece (2008).

(Problem/Game) Situation, part of the tutoring behavior framework, refers to the current, particular state of affairs within the inquiry environment; a particular point in the problem space. For example, the current formulas and values in an Excel spreadsheet problem, or the location of the Pacman, ghosts, power pellets, and score in Pacman.

(Problem/Game) situational feedback, part of the tutoring behavior framework, refers to changes in the situation which provide indirect feedback.

Stimuli, part of the motivation model, refer to sensations caused by a game or activity. Novel stimuli are thought to increase intrinsic pleasure. See Garris, Ahlers, & Driskell, (2002).

Synthesize, part of the deliberation framework, refers the process of combining one's current causal beliefs with new information to potentially create a new set of causal beliefs. Synthesis can be done with or without external representations.

Tutoring platform, refers to the cognitive model, learning challenges, and software that combined provide instruction.