International Journal of Artificial Intelligence in Education 19(2009) 401-424
IOS Press

Evaluating an Intelligent Tutoring System for Making Legal Arguments with Hypotheticals

Niels Pinkwart, *Department of Informatics, Clausthal University of Technology, Clausthal-Zellerfeld, Germany*
niels.pinkwart@tu-clausthal.de

Kevin Ashley, *Learning Research and Development Center, School of Law, & Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA*
ashley@pitt.edu

Collin Lynch, *Learning Research and Development Center & Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA*
collinl@cs.pitt.edu

Vincent Aleven, *Human-Computer Interaction Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA*
aleven@cs.cmu.edu

Abstract. Argumentation is a process that occurs often in ill-defined domains and that helps deal with the ill-definedness. Typically a notion of "correctness" for an argument in an ill-defined domain is impossible to define or verify formally because the underlying concepts are open-textured and the quality of the argument may be subject to discussion or even expert disagreement. Previous research has highlighted the advantages of graphical representations for learning argumentation skills. A number of intelligent tutoring systems have been built that support students in rendering arguments graphically, as they learn argumentation skills. The relative instructional benefits of graphical argument representations have not been reliably shown, however. In this paper we present a formative evaluation of LARGO (Legal ARgument Graph Observer), a system that enables law students graphically to represent examples of legal interpretation with hypotheticals they observe while reading texts of U.S. Supreme Court oral arguments. We hypothesized that, compared to a text-based alternative, LARGO's diagramming language geared toward depicting hypothetical reasoning processes, coupled with non-directive feedback, helps students better extract the important information from argument transcripts and better learn argumentation skills. A first pilot study, conducted with volunteer first-semester law students, provided support for the hypothesis. The system especially helped lower-aptitude students learn argumentation skills, and LARGO improved the reading skills of students as they studied expert arguments. A second study with LARGO was conducted as a mandatory part of a first-semester University law course. Although there were no differences in the learning outcomes of the two conditions, the second study showed some evidence that those students who engaged more with the argument diagrams through the advice did better than the text condition. One lesson learned from these two studies is that graphical representations in intelligent tutoring systems for the ill-defined domain of argumentation may still be better than text, but that engagement is essential.

Keywords. Legal argumentation, graphical representations, ill-defined domains, evaluation

INTRODUCTION

Much research on intelligent tutoring systems has focused on solving well-structured problems in domains such as mathematics and science. This work has been highly successful, resulting in systems that make a difference in many schools in the US and elsewhere (VanLehn, et al., 2005; Corbett, Koedinger, & Hadley, 2002; Mitrovic, Martin, & Suraweera, 2007). In order to move the field forward and make the technology widespread, however, it is important to branch out into other kinds of domains and problems as well.

The current work focuses on *ill-structured problem solving* involving argumentation. In domains like law, ethics, history, and public policy, argumentation is a fundamental tool for analyzing and reasoning about ill-structured problems. Well-structured problems usually state the goal and the applicable constraints a solution must address. Different people solving the same problem address the same or similar constraints, and the relevant community of practice agrees on what counts as a correct solution. By contrast, ill-structured problems often state the goal only incompletely and the applicable constraints not at all; the solver must refine the goal and infer the applicable constraints, and the way a solver frames the problem depends on his or her own knowledge, values, and interests. As a result, different solvers may frame the problem differently (Voss, 2006, p. 305-6). Ill-structured problems may have competing, even inconsistent, yet still reasonable solutions. Argumentation is essential in order to present, justify, and evaluate a solution. A proposed solution "usually is justified by verbal argument that indicates why the solution will work, and provides a rebuttal by attacking a particular constraint or barrier to the solution or by attempting to refute an anticipated opposing position." (Voss, 2006, p. 305-6). While the practice community may agree on what counts as a reasonable argument, a consensus about the correctness of a proposed solution is much less likely.

The current paper focuses on legal argumentation as a form of ill-structured problem-solving. In the legal domain, ill-structured problems are the norm. Legal problems often allow for alternative solution approaches, the correctness of which is a matter of degree, rather than a matter of applying objective criteria. For example, the fact that decisions by courts can be overturned on appeal is evidence of the ill-defined nature of (much) legal decision making: a single problem may have opposing reasonable answers. It is therefore essential that law students learn to "consider arguments counter to his or her argument and be able to refute them or to re-evaluate one's own position in reference to them.…" (Voss, & Means, 1991, p. 342).

Researchers aiming to develop systems that improve students' argumentation skills have been drawn to graphical representations for a number of reasons (Reed, Walton, & Macagno, 2007). First, from a cognitive perspective, graphical representations can reduce the students' cognitive load and reify important relationships. Thus, it is hypothesized, they facilitate reasoning about texts and the acquisition of interpretive skills (Ainsworth, 1999; Larkin & Simon, 1987). While the use of two simultaneous representations can increase cognitive load, the complementary strengths of a textual and graphical argument form can better guide students in their analysis. Proponents of argument diagrams also maintain that the diagrams can make the essential logical relations explicit while retaining formal validity. Second, intelligent tutoring systems can, in theory, provide feedback on graphical argument representations while finessing the fact that natural language processing remains difficult. A student-made graph provides the system with information about their thinking that, even if it does not rise to the level of complete understanding, can be leveraged to provide intelligent help (Paolucci, Suthers, & Weiner, 1996; Pinkwart, Aleven, Ashley, & Lynch, 2006a).

Work by Carr (2003) in the legal domain indicated that the production of argument diagrams can improve students' abilities to produce high quality arguments, and Schank and Ranney (1995) showed that the production of diagrams can improve students' argument coherence. Recent work by Harrell (2004) and by Easterday, Aleven, and Scheines (2007) has substantiated that argument diagrams can be useful learning tools. In summary, while controlled empirical studies are still rare, the current state of research suggests that computer-supported argument diagrams are a useful educational tool.

This paper presents a formative evaluation of LARGO ("Legal ARgument Graph Observer"), an intelligent tutoring system that helps beginning law students learn a form of argument that is widely used in law and other domains and is applicable to both formal and informal debate: skill in posing rules for deciding difficult cases, and testing (or "debugging") the proposed rule by generating hypothetical cases that challenge the rule and help explore the boundaries of its underlying rationale. The LARGO program presents students with examples of U.S. Supreme Court oral argument transcripts in which sophisticated legal minds thrust and parry with proposed tests, hypotheticals, and responses. In studying the examples, students see how competent professionals working under pressure deal with the ill-structure of legal problems presented by the need to decide the case. But students do more than just see good examples. In annotating and diagramming the examples of arguments, students reconstruct the arguments in terms of a model of hypothetical reasoning. The model is only partial; it reifies some essential structures but does not formalize the reasoning completely. By reconstructing the examples graphically, students impose enough structure on the examples as to make that argument structure visible, and they enable the program to prompt them to reflect on the arguments' merits and significance.

Developing a system that can generate meaningful feedback on graphical argument representations for ill-structured legal problems presents a special challenge for intelligent tutoring systems research. There are two sources of ill-structure that prevent the straightforward application of established intelligent tutoring systems techniques. First, deciding the legal problem is ill-structured, as discussed above. Underlying LARGO's new graphical language is a novel partial model of hypothetical reasoning, described further below, that captures some of the key elements in the argument transcripts, including tests, hypotheticals, and key relations among them. This model is focused on how attorneys defend or modify their proposed tests in response to the Justices' questioning by means of hypotheticals. Even if the reasoning captured in this model were the only decision-making process – there are others – given the facts of a legal dispute, there are many, many tests, hypotheticals, and responses that the decision makers can reasonably explore and justify. Second, for any given argument comprising some small number of tests, hypotheticals, and responses, even with LARGO's relatively modest set of graphical representation primitives, there are many plausible ways to diagram the argument. Beyond dealing with obvious extremes (e.g., an empty diagram), one simply cannot tell which diagrams are right or wrong, and there is no one ideal diagram.

LARGO is well adjusted to the doubly ill-defined nature of the domain: its help function does not assume a well-defined procedure for argumentation or argument analysis, nor does it insist on one standard diagram or assume that correctness of argument graphs can be determined objectively. Yet, it does present useful feedback to students: upon a student's request for feedback, it opportunistically identifies portions of the diagrams that appear problematic or complete enough in terms of their linkages to the text and their interconnections in terms of the model to be worth reflecting upon. It then prompts students to self-explain the argument transcript in terms of the model and the merits of the argument, and gives suggestions for possibly (but optionally) improving the graph. Even when

LARGO suggests that a modification of a student's argument graph may be appropriate, it does not actually force students to modify their graph.

The key research question we address in this paper is whether the help that LARGO provides students in imposing structure on a complex, real-world argumentation process is sufficient to help them learn the argumentation skills exemplified in the transcripts. More broadly, can an intelligent tutoring system operate in an ill-defined domain by supporting students as they graphically annotate argument transcripts, by providing a graphical argument language and on-demand feedback on student's argument diagrams?

We conducted two formative evaluation studies with first-semester law students (one with volunteers, one with mandatory participation as part of a first semester class) to test the hypothesis that LARGO's graphical representations and feedback help students learn better than a purely text-based tool that imitates the way law students would traditionally analyze texts. This article first describes the argument model underlying LARGO and the way in which the ITS helps learning and applying this model. Subsequently, the results of the two studies are presented.

LEGAL REASONING WITH TESTS AND HYPOTHETICALS

US Supreme Court Justices are famous for posing hypotheticals during oral arguments to evaluate an advocate's proposed rule for deciding the case before them. A proposed rule can be seen as a hypothesis, a tentative assumption made in order to draw out and test its normative, logical or empirical consequences. A hypothetical is an imagined situation that helps to test such a hypothesis; it is a tool for drawing out those consequences. The process of hypothetical reasoning typically unfolds as follows: an advocate proposes a decision rule, the Justices pose hypothetical cases in order to test the rules' consequences and how well the rule fits with past cases and underlying principles, and the advocate responds by analogizing the hypothetical to, or distinguishing it from the case, or by refining (or abandoning) the rule.

The hypothetical reasoning process is important; competent legal reasoners design hypothetical scenarios to integrate considerations at multiple levels: the facts of a case, its possible outcomes, the plausible rules justifying those outcomes, and how well they fit the underlying legal principles and past cases. For similar reasons, law professors pose hypotheticals in Socratic legal classroom discussions. The instructors have multiple pedagogical goals: to teach students something fundamental about the nature of legal reasoning - that attorneys and judges reason *about* legal rules, not just with them. The instructors also use the hypotheticals to explore the ins and outs of the legal rules in the subject matter domain of the course, for instance, contracts or constitutional law. The instructors seek to develop students' "critical legal imaginations" for hypothesizing scenarios in which the rule of law under consideration leads to unintended results or conflicts with deeply held norms. Students who learn this successfully can transfer their critical skills to the new legal domains they encounter in professional practice.

Although important, hypothetical reasoning is a difficult skill to learn in the legal classroom. Instructors engage as many students in classroom Socratic discussion as possible, but class sizes are frequently large. Students actively engage in arguments only sporadically; most of the time they only passively observe examples of argument exchanges between the professor and other students. The examples are oral and fleeting; there is little time for reflection and no detailed record beside the students' notes.

We illustrate the process of legal interpretation with hypotheticals, as employed within the LARGO ITS, with an extract from the oral argument in the case of *Kathy Keeton v. Hustler Magazine,* 465 U.S. 770 (1984). This case deals with one of the first technical legal concepts that U.S. law students encounter in the first year "Legal Process" course: personal jurisdiction, a court's power to require that a person or corporation appear in court and defend against a lawsuit. These cases often pit the principle that a state may redress wrongs committed within the state against the U.S. Constitutional principle of "due process" (i.e., minimum procedural safeguards against the arbitrary exercise of government power), especially when a court sitting in one state asserts power over a nonresident of that state. In the Keeton case, the plaintiff, Kathy Keeton, sued Hustler Magazine, an Ohio corporation with its principle place of business in California, in U.S. District Court in New Hampshire. She claimed that Hustler had libeled her in five articles published in the 70's. She was not a resident of New Hampshire and had almost no ties there. Hustler's contacts with New Hampshire included the monthly sale of up to 15,000 magazine issues there. At the time, New Hampshire was the only state in which Ms. Keeton was not barred under a state statute of limitations from making her claim.

The extract shown in Figure 1 illustrates key argument moves used during one of these sessions, modeled in terms of tests and hypotheticals as described above (see Ashley (2007) and Ashley, Lynch, Pinkwart, & Aleven (2008) for a detailed description of the argument model). The left column labels the different argument elements, such as proposed tests, hypotheticals, and ways of responding to hypotheticals, while the right contains the actual argument text. "Q:" indicates a Justice's question. Mr. Grutman represents Ms. Keeton. He begins by proposing a rule-like test for deciding the problem in a manner favorable to his client (line 14). Such proposals often include supportive reasons, such as that the proposed test explains past case decisions or is consistent with, or best reconciles, principles and policies underlying the law. Justices may respond by posing a hypothetical (lines 55, 57, 59, 126, 130, and 134), which may simply be a query about the test's meaning (line 55 and 57), or may underscore the test's overly broad scope (lines 59, 126, 130, and 134). The advocate has to rebut or otherwise reply to the challenge to maintain his argument's credibility. He may attempt to justify his proposed test by arguing that the supposedly disanalogous counterexample (i.e., the hypothetical) is really analogous to the current fact situation (cfs), in effect disputing that the proposed rule should yield a different result when applied to the counterexample than when applied to the cfs (as in lines 56 and 58). Or, he may distinguish the hypothetical from the cfs (as in lines 64, 66, and 127).

THE LARGO INTELLIGENT TUTORING SYSTEM

From the viewpoint of legal pedagogy, oral argument examples like that above are worth studying, but they are challenging materials to beginning law students. A program that engages students in reflecting upon such expert examples could help; it could bring the general argumentation principles to the forefront and at the same time require that students be active learners, not passive recipients as often occurs in larger law school classes.

→ **Proposed test of Mr. Grutman for Plaintiff Keeton**	14. GRUTMAN: The synthesis of those cases holds that where you have purposeful conduct by a defendant directed at the forum in question and out of which conduct the cause of action arises or is generated that satisfies the formula of those minimum contacts which substantial justice and reasonable fair play make it suitable that a defendant should be hailed into that court and be amenable to suit in that jurisdiction.
← **J.'s hypo**	55. Q: Would it apply in Alaska?
→ **Response: analogize cfs/hypo**	56. GRUTMAN: It would apply, Mr. Justice Marshall, wherever the magazine was circulated. It would apply in Honolulu if the publication were circulated there. It would apply theoretically and, I think, correctly wherever the magazine was circulated, however many copies were circulated
← **J.'s hypo**	57. Q: Just to clarify the point, that would be even if the plaintiff was totally unknown in the jurisdiction before the magazine was circulated?
→ **Response: analogize cfs/hypo**	58. GRUTMAN: I think that is correct, Mr. Stevens, so long as Alaska or Hawaii adheres, I believe, to the uniform and universal determination that the tort of libel is perpetrated wherever a defamatory falsehood is circulated. Wherever a third person reads about it, there is that harm
← **J.'s hypo**	59. Q: What if the publisher had no intention of ever selling any magazines in New Hampshire
	60. GRUTMAN: A very different case, Mr. Justice White.
	61, 63. Q: I know it is different, but how -- what would be - Would the result be different?
→ **Response: distinguish cfs/hypo**	64, 66. GRUTMAN: It might he different. It might be different, because in that case you could not say, as you do here, that you have purposeful conduct. There you have to look for other -- I think your phrase is affiliating circumstances, other connections, judicially cognizable ties --
← **J.'s hypo**	126. Q: Could she have filed 50 lawsuits?
→ **Response: distinguish cfs/hypo**	127. GRUTMAN: No, she could not, because the single publication rule requires that the plaintiff make an election of that jurisdiction in which she intends to make a claim not only for the harm that occurred in the jurisdiction where she properly brings suit, but for the harm that has occurred wherever the libel has been perpetrated.
← **J.'s hypo**	130. Q: Including Alaska and Hawaii?
	131. GRUTMAN: Including Alaska, Hawaii, Kampchatga and Tobago, wherever. I think --
← **J.'s hypo**	134. Q: Why don't you go to Guam while you're at it? (General laughter.)

Fig. 1. Examples of interpretive reasoning with hypotheticals in *Keeton v. Hustler*.

As noted, the LARGO system allows law students to graphically represent the dialectical pattern of hypothetical reasoning. Figure 2 shows an example graph based upon the Keeton excerpts in Figure 1. This graph was prepared by a naïve user for the purpose of illustration (since Keeton was part of the post-test for our study, the subjects did not use LARGO with this case). The left side of the screen shows the oral argument transcript. Below that is an advice button and a palette of the basic graphical elements used for markup. These elements include nodes representing proposed tests, hypotheticals, the current fact situation, and relations between them (test modification, distinction of and analogy with a hypothetical, hypothetical leading to test, and general relation). Graphs are constructed by

dragging these elements into the workspace on the right and filling in appropriate text. Students can also link the elements in their graph to passages in the transcript, using a text highlighting feature. The example diagram contains four test versions, six hypotheticals, two elements representing the facts of the case, and several relationships between these. For example, the hypothetical about a publisher who has no intention of publishing in New Hampshire is distinguished from the purposeful publication in the forum state (i.e., the state where the law suit is filed) in the current fact situation.

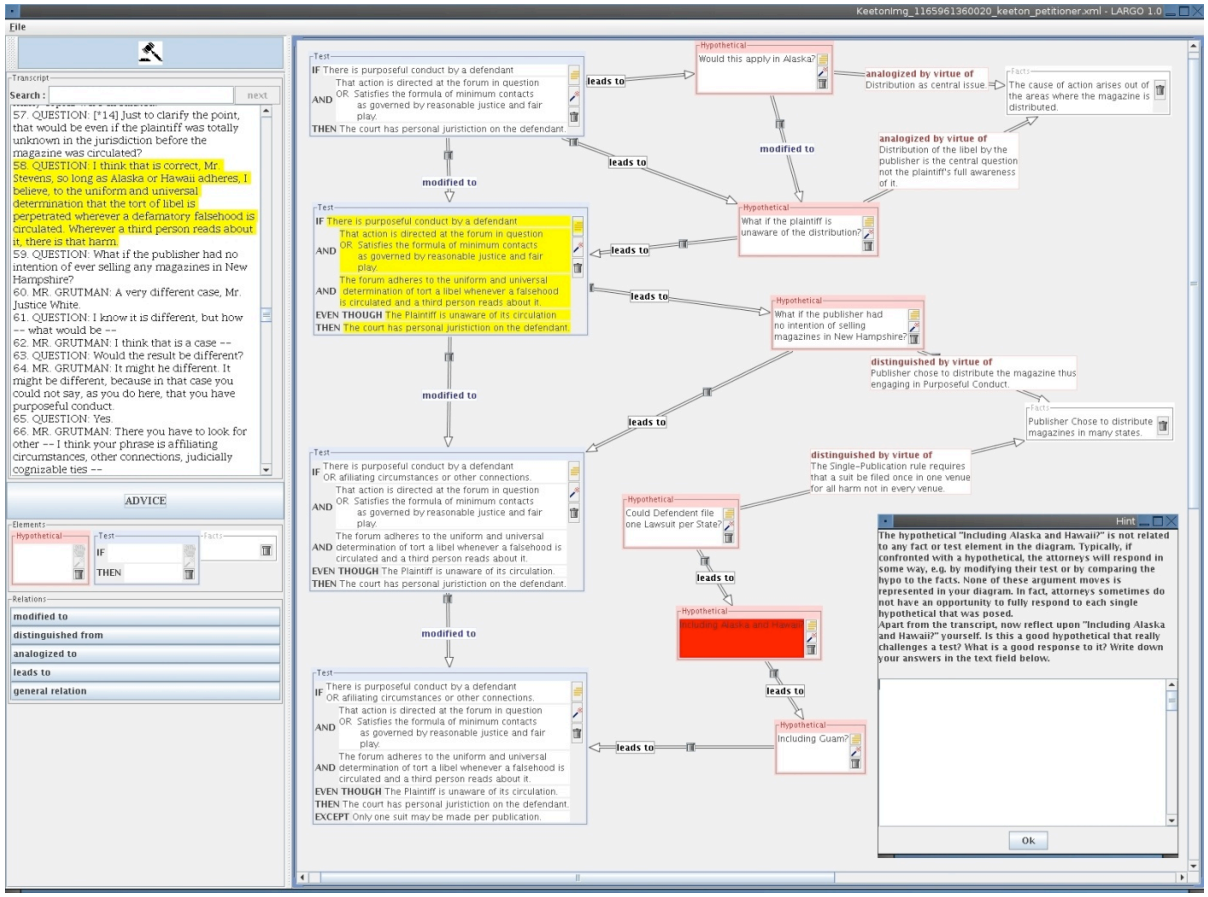

Fig. 2. LARGO Representation of Keeton Case Oral Argument.

LARGO guides students as they annotate and diagram the oral argument transcripts in terms of the described model of the hypothetical reasoning process. Based on an analysis of the argument diagrams (including its links to the transcript), LARGO provides two kinds of feedback on students' argument diagrams. A first type of feedback invites students to reflect on certain typical argument subpatterns found in their graphs. LARGO has a built-in graph-grammar engine that "knows" about several such patterns and is able to detect them (Pinkwart, Aleven, Ashley, & Lynch, 2006b; Pinkwart, Ashley, Lynch, & Aleven, 2008). LARGO also points to opportunities for extending these graphs, ranging from rather obvious omissions (e.g., key areas in the argument transcript not reflected in the argument diagram) to more intricate argument subpatterns that appear to be incomplete. The program suggests that a modification of the diagram may be in order, without forcing students to actually

implement the suggested changes. Given the ill-defined nature the legal domain (Lynch, Ashley, Aleven, & Pinkwart, 2006), one cannot always be certain that a diagnosed graph weakness represents an inaccurate rendition of the transcript, or how it should be "repaired." It may be that the particular line of argument is unusual (and it is difficult for system designers to foresee all such possibilities). Or it may be that the Justices abandoned a line of questioning before a standard argument pattern could be completed. Therefore, LARGO's feedback is typically couched as an invitation to reflect or as a self-explanation prompt. These types of prompts have proven effective as a metacognitive strategy (Chi, 2000) also in ill-defined domains (Schworm, & Renkl, 2002). For example, the hint box in Figure 3 (bottom right) prompts the student to think about the fact that one of the hypotheticals (about Alaska and Hawaii) is unconnected to any test or fact element in the diagram. If that was indeed the case in the transcript, then the diagram should reflect that (and thus is fine as it is), but if not, the invitation to reflect may lead the student to repair it. In either case it seems useful to invite the student to reflect on the role of this hypothetical.

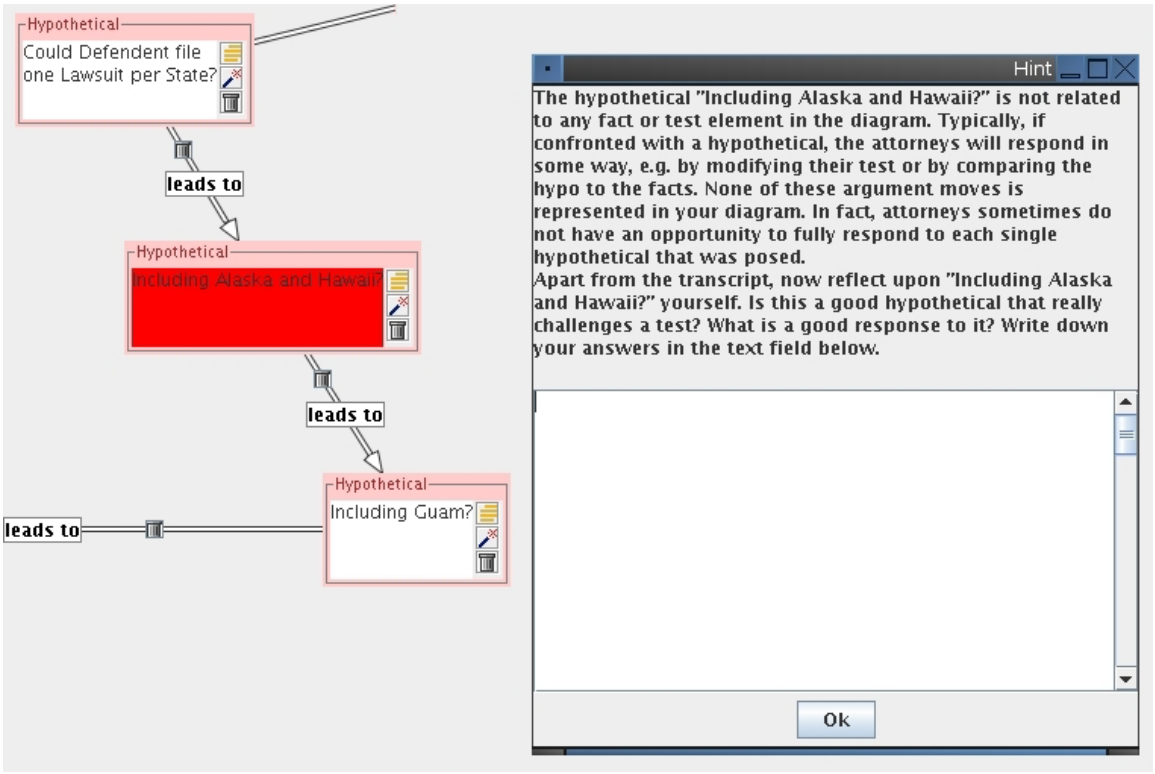

Fig. 3. LARGO feedback message.

A second type of feedback focuses on students' paraphrases of the decision rules that the attorneys propose. These paraphrases are short but rather complex pieces of text which often leave important elements of the tests implicit. Often, students must infer the details about a decision rule from the attorney's explanation for how certain hypotheticals should be decided, a process of distillation that is a key way of understanding the transcripts in greater depth. In LARGO, students rank each others' paraphrases by means of collaborative filtering techniques (Pinkwart, Aleven,

Ashley, & Lynch, 2006b). When students, after reading an attorney's proposed test in the transcript, enter a formulation of that test into a test node in their diagram, they are prompted to rate their formulation of that test against others produced by their peers or the professor. This information enables LARGO to derive a ranking of the formulations of a given test by all students (without applying natural language techniques) and to ask students with presumably poor formulations to revise the test descriptions in their diagrams.

In summary, while LARGO typically has a lot to say about a student's argument diagram (and the more elaborate the graph, the more feedback it can typically provide), it does so only on demand, and, in contrast to the more typical intelligent tutoring systems, it accommodates the ill-defined nature of the domain by never forcing students to adopt its viewpoint.

Many of the aspects of LARGO have not been studied in the ITS and legal argumentation fields before. This includes the students' tasks of analyzing arguments in order to train their argumentation skills, the idea of using U.S. Supreme Court oral argument transcripts as expert examples of legal reasoning, the graph grammar and collaborative filtering based analysis components of LARGO, and the way that on-demand adaptive feedback is given as self explanation prompts. To investigate whether this combination of novel ideas and approaches indeed helps law students learn and to find out if parts of this system design may need to be revised, we conducted an extended formative evaluation of the system.

STUDY 1

As a first part of our formative evaluation of LARGO, we conducted a pilot study comparing LARGO's graphical representations and advice with a purely text-based alternative. The study was a first test of the hypothesis that, compared to standard studying techniques, a special-purpose diagramming language geared toward depicting hypothetical reasoning processes, coupled with non-directive feedback, helps students better extract the important information from argument transcripts and better learn argumentation skills. The text-based Control condition was designed to reflect the way that law students usually analyze given texts – they highlight text passages in the material and take notes. Figure 4 shows a screenshot of the tool used in the Control condition. The design of this study thus compares the newly proposed tool LARGO to "traditional tools" (or a computer approximation thereof) used by students. Given that the conditions differed both with respect to the representations used (text v. diagrams) and with respect to the availability of feedback, the study could be said to investigate the diagonal of a 2x2 design. With the limited number of available participants for the study, it was reasonable first to investigate if LARGO helps students learn at all, and then potentially use these results to improve the system and to study in follow-up experiments which factors of LARGO are essential.

The study was conducted in concert with four sections of the 2006 first year Legal Process course at the University of Pittsburgh's School of Law. With the professors' permission, we invited students to volunteer for the study. The core cases examined in the study all centered on questions of personal jurisdiction. This topic was part of their coursework. The students were assigned randomly to the conditions, balanced in terms of LSAT scores (the Law School Admissions Test, which is a frequently used predictor of success in law schools). Students were paid $80 for their participation.

The study involved four sessions of 2 hours each over a four week period. The first session included a pre-test, a short introduction to the software and instructions about the model of

hypothetical reasoning, including an extended example. In the second and third sessions, the students worked with extracts of the oral arguments from two personal jurisdiction cases. In the Experimental condition, students represented them graphically using LARGO with the help of the feedback mechanisms. In the Control condition, students were instructed to highlight relevant passages and take notes using the text based tool. Session four consisted only of the post-test. No argument representation tools were used during this session.

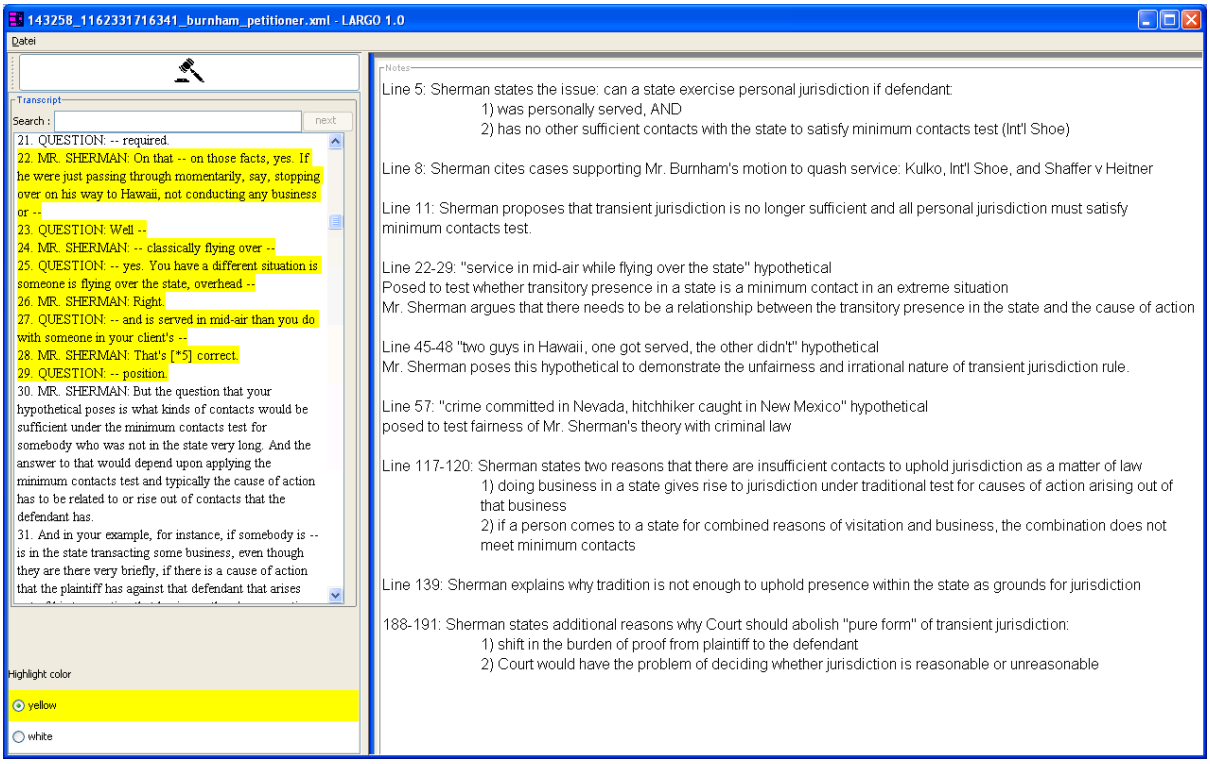

Fig. 4. Screenshot of Control condition (text tool).

The pre- and post-tests, designed to assess students' argument and hypothetical reasoning skills, comprised five types of multiple choice questions: a) legal argument related questions of a type considered for inclusion in the LSAT (Reese, & Cotter, 1994); b) generic questions about the use of hypotheticals in legal argument; c) "tennis club" questions that explored the use of hypotheticals for argument in a non-legal, intuitive everyday domain about the policies of an imaginary tennis club; d) the domain of personal jurisdiction; e) near transfer argumentation questions involving selecting proposed tests, hypotheticals, and responses in a new personal jurisdiction case, namely Keeton; f) far transfer argumentation questions similar to those in personal jurisdiction but drawn from a new legal domain (copyright law) with which first year law students are not likely to be familiar. The first three problem types appeared on both pre- and post-test; the last two appeared only on the post-test. Both tests were created with the CTAT tools and delivered via the web (Aleven, Sewall, McLaren, & Koedinger, 2006). The test items were not formally checked for validity and reliability. However, they have face validity, as reported by an experienced law school professor and some advanced graduate law students who took the test. The validity of the items is also supported by our study results that

students with higher LSAT scores (as mentioned, a national standardized test used in the US to evaluate students' aptitude for law school) scored significantly higher on the test (cf. next section) and that a sample of 17 more advanced third year law students who went through the same procedure as the first year students in our study 2 (including the same tests) performed significantly ($p<.05$) better on the post-test (m=.71, sd=.05 for third year vs. m=.59, sd=.09 for first year) and on the pre-test (m=.63, sd=.11 for third year vs. m=.58, sd=.11 for first year) than the beginners. See Lynch, Ashley, Pinkwart, and Aleven (2008) for a description of the study with the more advanced students.

RESULTS OF STUDY 1

Of the 38 students who began the study, 28 completed it, 15 in the Experimental condition and 13 in the Control condition. These students (15 female, 13 male) had LSAT scores between 158 and 165 (m=160.9, sd=1.8) – as a baseline, the average score of students accepted for the University of Pittsburgh's law school is 159, with 25th and 75th quartiles at 158 and 161, respectively. While they had a maximum of two hours time to work on each of the training cases, their average time per case was 77 minutes (sd=25). There was no significant training time difference between the two conditions.

For each participant, we first computed a single overall pre-test and post-test score which included all multiple choice survey items. We also computed subscores for each of the five specific question types described above. No significant difference on overall or subscores existed between the two conditions on the pre-test. Table 1 contains the post-test scores for the two conditions on a scale from 0 to 1. As the data shows, the averages of the post-test scores for the Experimental subjects were higher than the Control subjects' scores with respect to overall scores and also when looking at four out of the six subscores. However, the sample size of this first study was very small, and the differences were not statistically significant ($F(1,26)=.83$, $p>.3$, for the overall scores).

Table 1
Results of Study 1 (N=28)

mean (sd) of post-test score	Control	LARGO
All	.55 (.07)	.58 (.07)
LSAT Questions	.56 (.20)	.48 (.18)
Generic Argumentation	.67 (.14)	.64 (.29)
Everyday Argumentation	.79 (.13)	.85 (.12)
Personal Jurisdiction	.38 (.27)	.42 (.20)
Near Transfer	.52 (.07)	.56 (.10)
Far Transfer	.53 (.13)	.57 (.13)

We then divided up students by LSAT score, creating a "Low" group containing 10 students, a "Medium" group with 9, and a "High" group with 8 (the group sizes vary slightly to ensure that all students with the same score are in the same group; all students in the Medium group had the same LSAT score). One student did not have an LSAT score and was not included. The results of these three groups differed considerably ($F(2,24)=3.79$, $p<.05$, for the overall score, similar results for most subscores). The students in the Low group (average post-test score .54) scored significantly lower ($p<.01$) than those in the High group who had an average of .62. The Medium group scored in between the two (average .55). Within the Medium and High groups, the Control and Experimental conditions

were not equally distributed (Medium: 8 Experimental, 1 Control; High: 2 Experimental, 6 Control), so that meaningful results of an in-depth statistical analysis could not be expected and we do not report further statistics here. The Low group on the other hand was balanced, with 5 Experimental and 5 Control subjects. We hypothesized that lower-LSAT students would do better with LARGO than with the text based annotation tool. This hypothesis was confirmed with respect to all item types but the LSAT questions (see Table 2). In particular, we found a significant condition effect for the post-test near transfer questions. A 1-sided t-test confirmed this hypothesis, showing an effect size of 2.25 ($p<.05$). There was no pre-test difference between the Experimental and Control subgroups in the Low group ($p>.8$ in the overall pre-test score – the pre-test did not include specific Keeton questions).

Table 2
Results of lower LSAT students in Study 1 (N=10)

mean (sd) of post-test score	Control	LARGO
All	.52 (.06)	.57 (.03)
LSAT Questions	.47 (.18)	.40 (.09)
Generic Argumentation	.67 (.00)	.73 (.28)
Everyday Argumentation	.74 (.15)	.82 (.12)
Personal Jurisdiction	.33 (.24)	.40 (.15)
Near Transfer	.52 (.04)*	.61 (.07)*
Far Transfer	.46 (.15)	.49 (.10)

* $p<.05$

Another way of classifying the items in the questionnaires is to group them by the aspect of the argument model they relate to most: 1) tests, 2) hypotheticals, 3) relations between test and hypotheticals, 4) responses to hypotheticals, 5) legal issues, or 6) legal policies. A post-hoc analysis of the study results based on this grouping of items revealed some interesting findings: in general, students in the Low group benefited from LARGO more than students in the other two groups. Specifically, lower aptitude students using LARGO did significantly better in the post-test on questions about legal issues than their peers in the Experimental group ($p<.05$, see table 3); and students in the Low and Medium groups benefited from LARGO and did better on post-test questions that asked them to evaluate a hypothetical with respect to a given test. For the combined Low+Medium group (i.e., all but the top third of the participants), the difference was significant ($F(1,17)=7.41$, $d=1.00$, $p<.05$, 1-sided), but not for the whole group or for the Low group alone.

Table 3
Results of Study 1 by argument model aspect

mean (sd) of post-test score	All students (N=28)		Low-LSAT students (N=10)	
	Control	LARGO	Control	LARGO
Tests	.34 (.21)	.35 (.13)	.28 (.18)	.39 (.11)
Hypotheticals	.64 (.08)	.62 (.09)	.65 (.06)	.60 (.05)
Relations Tests / Hypotheticals	.58 (.13)	.64 (.09)	.50 (.14)	.63 (.06)
Responses to Hypotheticals	.49 (.12)	.56 (.18)	.58 (.08)	.56 (.13)
Legal Issues	.50 (.35)	.73 (.26)	.40 (.22)*	.80 (.27)*
Legal Policies	.58 (.28)	.53 (.35)	.50 (.00)	.50 (.50)

* $p<.05$

Figures 5 and 6 show examples of "hypothetical evaluation with respect to test" and "legal issues" questions where the Experimental subjects outperformed the Control subjects.

Assume that Mr. Grutman's proposed test is as follows: If the state long-arm statute is satisfied and defendant has engaged in purposeful conduct directed at the forum state out of which conduct the cause of action arises, and that conduct satisfies the minimum contacts under which substantial justice and fair play make it reasonable to hail defendant into court there, then the forum has personal jurisdiction over the defendant for that cause of action.

The following hypotheticals either were or could have been posed in the oral argument. Each of them is followed by four explanations why the hypothetical is or is not problematic for Mr. Grutman's proposed test. For each hypothetical, please check ALL of the explanations that are plausible.

"Just to clarify the point, that would be even if the plaintiff was totally unknown in the jurisdiction before the magazine was circulated?" [i.e., suppose the plaintiff was totally unknown in the state before the magazine was circulated. Would personal jurisdiction over Hustler Magazine lie in that state?]

- o The hypothetical is problematic for Mr. Grutman's proposed test. The decision rule applies by its terms, but arguably the publisher should not be subject to personal jurisdiction in the state under those circumstances.
- o The hypothetical is not problematic for Mr. Grutman's proposed test. The decision rule applies by its terms, and the publisher should be subject to personal jurisdiction in the state under those circumstances.
- o The hypothetical is problematic for Mr. Grutman's proposed test. The decision rule does not apply by its terms, but arguably the publisher should be subject to personal jurisdiction in the state under those circumstances.
- o The hypothetical is problematic for Mr. Grutman's proposed test. The decision rule applies by its terms, but publishers would then be subject to personal jurisdiction even in a state where defendant suffered no injury.

Fig. 5. Example post-test question: relations test / hypotheticals.

What legal issue concerning Keeton's appeal in Kathy Keeton v. Hustler Magazine, et al. did the Justices address in the oral argument excerpt? Select the best answer below.

- o In determining whether courts in a state may exercise personal jurisdiction over a defendant, should the court consider whether the statutes of limitations have run out in other states where the action could otherwise have been brought?
- o May courts in a state exercise personal jurisdiction over an out-of-state publisher in a libel action by an out-of-state plaintiff where the publisher sold 10,000 to 15,000 magazines in the state on a monthly basis?
- o Are statutes of limitations state procedural law or substantive law for purposes of determining whether the courts of the state have personal jurisdiction over an out-of-state publisher?
- o Since a plaintiff in a libel action may prove damages occurring in other states under the single publication rule, should a court have personal jurisdiction over a defendant in any state where the damage occurred?

Fig. 6. Example post-test question: legal issues.

These results support our research hypothesis (although perhaps fall somewhat short of decisively confirming it). For the Low group, the use of LARGO with its graphical argument representation and feedback in the form of tailored self-explanation prompts led to significantly better learning of legal argumentation skills than the use of traditional note-taking techniques, as measured in a near transfer problem which involved argumentation questions about a novel case in the same legal domain as those studied in the study. For the far transfer problem (a novel case in different legal domain), this effect was not found. We were also intrigued by the findings that LARGO helped lower aptitude students learn legal issues, and that Low and Medium Experimental subjects apparently learned more about evaluating hypotheticals with respect to tests than their Control counterparts. As the example in Figure 5 illustrates, this skill is central to what LARGO is designed to teach: the essential relationship between tests and hypotheticals in legal argument.

One important question is why a significant difference was found on this particular question type and not on the other main items that were related to the argument model (tests, hypotheticals, responses to hypotheticals). One possible explanation is that LARGO's graphical language distills the essence of the oral argument visually, explicitly identifying the relations between tests and hypotheticals (cf. Figure 2). Our data suggests that less skilled students benefited from creating and reflecting on these diagrams (with the help of LARGO's feedback), whereas more skilled students may have been able to understand the complex relations without aid. One can argue that for the other items that are related to the argument model, the specific graph structure or advice features that LARGO employs are not sufficient to differentiate it from purely text-based annotation tools. The student's ability to formulate a good test might not be supported to a great extent by a graphical representation format or prompts LARGO offers. However, as the near transfer effect for the Low group indicates, the less skilled students did benefit from LARGO also on a general level.

In summary, our analysis of the results of this first study suggests that the LARGO ITS can be a valuable tool for those learners who do not (yet) have the ability to learn argumentation skills from independent study of argument transcripts. This group seems to benefit from the scaffold that the diagrams and the feedback offer. For the more advanced/skilled students, LARGO did not prove to be significantly better (but also not worse) than traditional learning resources such as a notepad and a highlighter.

However, these findings have to be taken with care. The results for subgroups (like the lower ability students) are based on very few students, and the Medium group additionally was highly imbalanced. In addition, the set of participants in the study might not be fully representative of a typical law school class. Since participation in the first study was voluntary, the students were self-selected for their interest in the curriculum, the ITS, and the remuneration. Many of the study participants expressed an interest in the system, making it apparent that they were among the more inquisitive members of their class. For these reasons, we decided that a second formative evaluation study in a real course setting was necessary to further examine and substantiate the findings with more and non-voluntary participants, including more lower-LSAT students. Such students may need extra help, and yet may have been less likely to volunteer for our first study (an apparently time-consuming "extra" educational activity) regardless of financial remuneration.

STUDY 2

We developed a curriculum covering three personal jurisdiction cases in LARGO integrated into one section of the 2007 first year Legal Process course at the University of Pittsburgh's School of Law. All 85 students in the section were required to complete the curriculum. These students (43 female, 42 male) had LSAT scores between 146 and 174 (m=159.3, sd=3.8). They were not paid but were given coffee gift cards as a token of appreciation. Students were assigned randomly to one of three course sections so we have every reason to believe that this group is representative of their peers. The curriculum was integrated into the class as preparation for a graded writing assignment on personal jurisdiction, counting for 10% of their grade. The participation in the study itself was mandatory but did not count for the grade.

Study 2 addressed the same hypothesis as study 1. The students were randomly assigned to two study conditions, balanced by LSAT scores. As in the first study, the Experimental group used a graphical version of LARGO that supported diagram creation and gave advice. The Control group made use of the text version that offered no feedback. The curriculum consisted of six weekly two hour sessions. In the first week, the students took a multiple choice pre-test, received instruction about the model of reasoning with hypotheticals and an extended example, and were introduced to the software (as was done in the first study). During the second to fourth week, they read background material on personal jurisdiction cases and annotated both sides (petitioner and respondent) of the transcript in LARGO or the text tool. They then answered two written questions about it without their diagrams or notes. Two of the three cases were part of the 2006 study – i.e., the training time was increased by 2 hours (or: 50 percent). During week five, they took a post-test consisting of multiple choice and free answer questions. Finally, we offered a debriefing session to show students in each condition the version the other had worked with. We sought to eliminate any residual post-test differences between conditions prior to the writing exercise or the course examination.

The pre- and post-tests used in this study were very similar to the test used in study 1. Both tests (pre and post) again contained multiple choice questions about: a) legal argument-related LSAT questions; b) generic questions about the use of hypotheticals in legal argument; c) "tennis club" everyday argumentation questions; and d) domain questions about personal jurisdiction. Questions of types a) to d) were counterbalanced between pre- and post-test. The post-test additionally contained e) analysis and free-text questions regarding a novel case (the near transfer questions from study 1) as well as f) factual recall questions; and g) interpretation questions, both regarding the transcripts studied during training. Since there were no differences between conditions in the far transfer questions in the first study, we did not include these (relatively time consuming) questions in the post-test this time. As in the first study, we also grouped the items with respect to the aspect of the argument model that they most related to.

RESULTS OF STUDY 2

All 85 students completed the study. While they had a maximum of two hours' time to work on each of the training cases, their average time per case was 55.8 minutes (sd=13.3). There was no significant training time difference between the two conditions.

We excluded a total of 15 students from the analysis. Four candidly told us that they were not working and explicitly entered off-topic responses in the post-test. Two others completed the post-test in less than 30 minutes, less time than is needed to merely read the materials (approx. 50 minutes). The remaining nine spent less than 30 minutes on one or more of the training cases, less time than it takes an expert to work through the material (approx 45 minutes). It is therefore highly unlikely that they put considerable effort into their task. The analyses below are based upon the remaining 70 students (36 Text, 34 LARGO).

Table 4 contains the mean scores (on a [0;1] scale) and standard deviations of the case-specific post-test questions. Table 5 shows the pre-post gains for counterbalanced items shared between the tests. Both tables show the results for all 70 students as well as the sub-results for the 27 low-LSAT students whose LSAT scores were below the median of 159. For this group, study 1 showed a positive effect of LARGO as compared to the text tool.

Table 4
Study 2 results for post-test only items

mean (sd) of post-test score	All students (N=70)		Low-LSAT students (N=27)	
	Control	LARGO	Control	LARGO
All	.63 (.09)	.64 (.09)	.64 (.08)	.61 (.11)
Near Transfer	.40 (.13)	.39 (.11)	.45 (.10)	.49 (.11)
Case Interpretation	.46 (.11)	.48 (.10)	.45 (.10)	.49 (.11)
Case Recall	.71 (.10)	.73 (.12)	.73 (.09)	.67 (.14)
Tests	.75 (.18)	.79 (.15)	.75 (.17)	.76 (.21)
Hypotheticals	.71 (.12)	.71 (.14)	.72 (.13)	.64 (.14)
Relations Tests / Hypotheticals	.48 (.11)	.50 (.11)	.49 (.11)	.48 (.16)
Responses to Hypotheticals	.44 (.23)	.45 (.24)	.40 (.32)	.49 (.28)
Legal Issues	.39 (.49)	.35 (.48)	.50 (.52)	.38 (.51)
Legal Policies	.36 (.49)	.29 (.46)	.50 (.52)	.23 (.44)

As table 4 shows, there were no significant differences between the two conditions with respect to post-test only test items – neither overall nor for the lower LSAT subjects.

Table 5
Study 2 results for items counterbalanced between pre- and post-tests.

mean (sd) of gain score	All students (N=70)		Low-LSAT students (N=27)	
	Control	LARGO	Control	LARGO
All	-.01 (.16)	-.04 (.18)	-.01 (.13)	-.08 (.19)
LSAT Questions	-.03 (.23)	-.02 (.24)	-.02 (.20)	-.06 (.25)
Generic Argumentation	-.01 (.31)	-.01 (.27)	-.02 (.28)	-.03 (.25)
Everyday Argumentation	.01 (.34)	-.05 (.36)	.09 (32)*	-.19 (.38)*
Personal Jurisdiction	.07 (.40)*	-.13 (.42)*	.00 (.35)	-.21 (.32)
Tests	-.17 (.65)	-.18 (.52)	-.36 (.63)	-.15 (.55)
Hypotheticals	.08 (.53)	.00 (.49)	.21 (.42)	.00 (.41)
Relations Tests / Hypotheticals	-.01 (.22)	.01 (.30)	-.03 (.24)	-.07 (.40)
Responses to Hypotheticals	.06 (.39)	.01 (.34)	.14 (.36)	-.12 (.36)

* $p<.05$

For the shared question types (see table 5), the Control group gained significantly more than the LARGO group on the personal jurisdiction items ($F(1,68) = 4.250$; $p<.05$) – items that LARGO does not directly teach (they focus on domain knowledge rather than argumentation skills). For the low LSAT students, the Control group gained significantly more than the LARGO group in the "everyday argumentation with hypotheticals" questions ($F(1,25) = 4.313$; $p<.05$). No other significant differences were found. The data further shows that neither group seemed to benefit from the study. A repeated measures analysis reveals that the only significant difference between the pre-test and post-test scores is a pre-post drop for the Low-LSAT LARGO students ($F(1,10)= 5.333$; $p<.05$) in the "personal jurisdiction" domain questions.

This apparently contradicts our results of study 1 where the Low-LSAT LARGO students outperformed their Control peers on several important question types. While the mode of participation differed (voluntary vs. mandatory), the two studies were very similar in many aspects: the same tools have been used with the same cases, a similar population (first semester law students at the same University) and very similar instructions. The training time in study 2 was longer than in study 1 (6 hours vs. 4 hours), and study 2 involved more participants than study 1 (70 vs. 28). These two differences between the studies do not explain the contradicting results – we would have expected even stronger results in study 2 based on them.

So why were the results of the two studies different, suggesting that argument diagrams in combination with feedback helped some students acquire certain argumentation skills in the first study but not in the second? One possible explanation is that the sample size in study 1 was very low (10 students in the Low group) and that the results were due to chance. In order to investigate whether this is the only possible explanation or whether there is any evidence for alternative explanations, we undertook a more detailed analysis of the data files and log files produced during the studies. These analyses, presented in the next two sections, aim at understanding the student's use of the two key features of LARGO: the note taking functions and the intelligent feedback.

STUDENT'S NOTE TAKING

We first investigated whether there were any differences in the ways that students took notes, across conditions and across studies. In particular, we were interested in finding out if students using LARGO were better at finding and attending to important portions of the transcript text than students who use the text tool. This is plausible, since LARGO gives feedback that points students to these important passages.

We began our analysis by determining how much of the students' work was *relevant* (that is, forwarded the goals of their analysis) and how much of it was not. Our particular focus was on the students' identification of the relevant tests and hypotheticals within the transcript.

In preparation for the analysis, an expert legal instructor marked up the redacted transcripts of each argument and identified a set of important tests and hypotheticals in each oral argument. This resulted in a list of 33 regions over the two cases that were used in both studies. Sixteen of these regions (the "Core" set) were encoded into LARGO for use in providing hints. The remaining 17 were reserved for this analysis and designated as the "Test" set.

In order to effectively compare the two conditions (text vs. LARGO), we defined a standard baseline unit of student work, a note, as a single atomic reference or notation made by the students. For LARGO students, a note is a single graph node or relation – for example, the graph shown in

figure 2 contains 29 notes. As stated before, the test and hypothetical nodes may be linked to the text transcript. A note is relevant if it is linked to one of the Core or Test set transcript locations and is of the correct type.

Including relational links and fact nodes in this calculation tends to penalize the LARGO subjects for making those kinds of notes as they cannot increase the success measures, only decrease them: Since the relations and fact nodes could not be linked to the transcript, they count as "non relevant" notes. We opted to include them for three reasons: (1) the students' task was to mark up the transcript including relations and discounting that effort would skew the counting toward minimal graphs; (2) the relationship structure has value and should be a part of any reasonable assessment; (3) dropping the edges unilaterally from the LARGO condition would bias the results in their favor as no viable standard was available for discounting text notes in the same way.

For the text condition, a note is defined as a single paragraph entry that may be accompanied by a highlight. Such a note is relevant if the text explicitly references some key transcript portion by line number or via a highlight and if it explicitly identifies the type of the location in text. Figure 4 contains 9 textual notes, and the 4th, 5th, and 6th all specify a type. We defined note in this way to ensure that the text and LARGO subjects employed roughly the same amount of cognitive effort when making each note. For the text subjects, we approximated the total number of notes by using the number of highlights or text notes (whichever was larger). Thus we kept the count linked to distinct note-taking acts. While this may undercount slightly, we think that it is a viable choice.

We analyzed the quality of student's note taking based on three success measures (efficiency, precision, and recall) that are commonly employed in machine learning applications. These success measures reflect the extent to which the student did or did not focus on the key elements. Recall is defined as the number of relevant notes in a student diagram divided by the total number of important passages in the transcript (i.e., how good was the student in finding the important information in the text?). Precision is the number of relevant notes in a student diagram divided by the total number of notes in the diagram (i.e., how many of the student's notes were about important things?). Efficiency is defined as the number of relevant elements located divided by the time on task. Recall and precision are measured on a scale from 0 to 1 (with 1 being a perfect result), while a higher value for efficiency indicates a better performance of the student. These definitions vary slightly from those typically used in machine learning but we find them more appropriate here. We calculated each measurement with respect to the Core and Test sets.

Table 6 and Figure 7 show an overall comparison between the text and LARGO groups in study 1 and the LARGO group in study 2 (due to the large N in the second study, a manual coding of the notes created by the text condition subjects was not conducted).

Table 6
Study comparisons: efficiency, precision and recall of student's note taking

Mean (sd)	Core set			Test set		
	Study 1 Text	Study 1 LARGO	Study 2 LARGO	Study 1 Text	Study 1 LARGO	Study 2 LARGO
Efficiency	.07 (.04)	.11 (.03)	.02 (.02)	.05 (.03)	.04 (.02)	.02 (.02)
Precision	.08 (.04)	.23 (.07)	.06 (.05)	.06 (.04)	.07 (.04)	.04 (.04)
Recall	.39 (.20)	1.00 (.00)	.15 (.11)	.26 (.14)	.32 (.15)	.12 (.11)

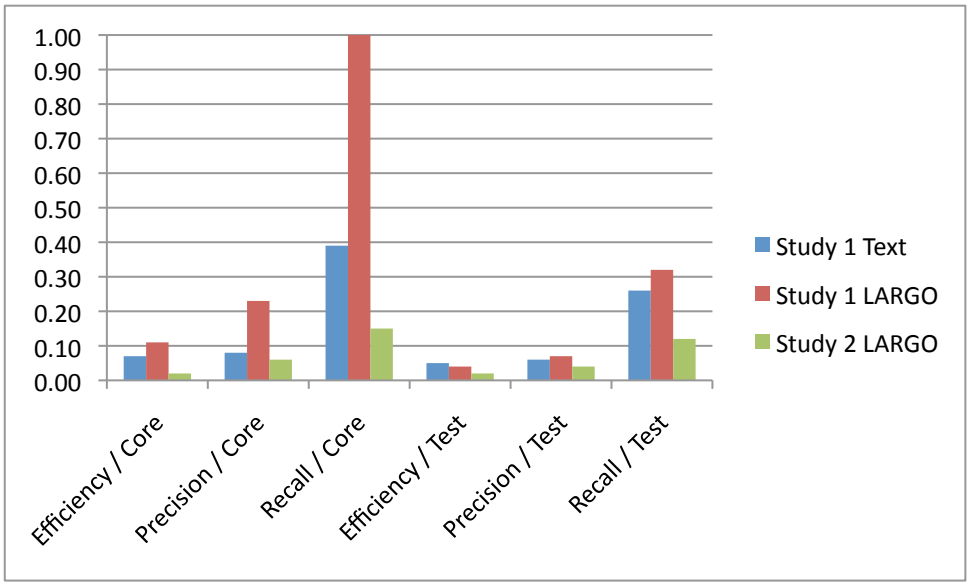

Fig. 7. Study comparisons: efficiency, precision and recall of student's note taking.

As the table shows, the LARGO condition in study 1 overall outperformed the Control condition in terms of efficiency, precision and recall on the Core set on which LARGO gave advice (which also holds when considering only low-LSAT students, cf. Lynch, Ashley, Pinkwart, & Aleven, 2007). This dominance did not hold when measuring against the Test set, consisting of elements that LARGO did not point students to. There, the two conditions were equal overall. These differences between conditions and between Core and Test suggest that the LARGO advice was effective in supporting the students' note-taking in study 1, even though it did not explicitly state the missing tests or hypotheticals but only pointed the students to a region of interest.

A comparison of the performance results of the LARGO subjects between the two studies shows that in all three performance measures (precision, recall and efficiency), LARGO students in the second study did significantly worse than their counterparts in the first study. This is true for both the Core set (which the system gave hints on) and for the Test set. On all note-taking performance measures, the LARGO students in study 2 did not only worse than their counterparts in study 1, but even worse than the Control subjects in study 1 who did not get any support.

STUDENT'S USAGE OF ADVICE FUNCTIONS

Having found some considerable differences of the quality of student's note taking between the studies and also between conditions in the first study, our next step in the data analysis was to look at the students' use of the second main LARGO function, the advice. We counted how often the students in the LARGO conditions of the two studies made use of the advice functions of the system.

On average (across all sessions of the study and all students in the Experimental condition), the advice button was pressed 10.1 times per transcript (i.e., approx. per hour) during the first study. Students of all aptitudes frequently requested advice (Low 12.3; Medium 6.2; High 17.9). In 75% of these cases, students selected one of the three short hint titles that LARGO presented in response to

their hint request and read through the detailed feedback related to the selected hint title. The use of the advice did not decrease over time. In the later sessions, the average number of help requests was even higher than in the earlier sessions (12.2 and 8.6 in the last two transcripts vs. 7.3 and 9.8 in the first two transcripts), which we consider as evidence that the students must have considered the advice to be valuable. The data of the second study shows a different picture. Table 7 shows the results of this comparison: The students in the first study made far more use of the advice functions than the non-volunteers in the second study. Moreover, the advice usage of the non-volunteers dropped over time. During the last session, on average only 0.6 advice requests were made per case (1.6 during the first case).

Table 7
Advice usage and diagram complexity

mean (sd)	first study (N=15)	Second study (N=34)
Clicks on Advice button (shows 3 hints) per transcript	10.1 (10.8)	1.8 (3.9)
Selection of one of the 3 shown hints per transcript	7.6 (8.2)	1.2 (2.2)
Advice usage by case over time	increasing from 7.1 to 8.1	decreasing: 1.6, then 1.3, then 0.6

We next conducted a correlation analysis in order to further investigate if within the second study, a higher number of advice requests correlates with higher post-test or gain scores. Table 8 contains these results.

Table 8
Correlations between advice requests in LARGO and test scores in study 2

Pearson correlations	All (N=34)			Low-LSAT (N=13)		
	Pre	Post	Gain	Pre	Post	Gain
Case Interpretation	-	.03	-	-	.15	-
Case Recall	-	-.05	-	-	.02	-
LSAT Questions	-.07	.02	.06	-.11	.30	.24
Generic argumentation	-.06	-.19	-.18	.06	-.14	-.21
Everyday argumentation	-.06	.34 *	.33	.07	.46	.29
Personal Jurisdiction	-.09	.21	.16	.04	.46	.30
Tests	.05	.06	-.03	-.07	.11	.16
Hypotheticals	-.02	-.19	-.04	.24	-.18	-.28
Relations Tests / Hypotheticals	-.09	-.20	-.17	.28	-.29	-.37
Responses to Hypotheticals	-.15	.29	.33	-.16	.54	.61 *

*: $p<.05$

As it shows, the pre-test scores are not significantly correlated with advice usage: better students did not tend to use help more or less often. For the post test scores, there are two statistically significant results: advice usage is positively correlated to the post-test score for everyday argumentation items for all subjects, and for Low-LSAT students the advice usage is also highly positively correlated to pre-post gains in items about responses to hypotheticals. The advice given by the system, apparently, helped these students to better understand how one can react to a hypothetical during (legal or everyday) argument. These strategies are indeed contained in the feedback messages LARGO provides.

DISCUSSION

The results of the post-hoc analyses of help usage and note taking suggest an explanation for the result that LARGO was apparently helpful for lower tier students in study 1 but not in study 2. The voluntary participants in study 1 were much better than the non-voluntary participants in study 2 in their note taking, covering more of the highly relevant parts and less of the less relevant parts in shorter time. This can – at least partially – explain the post-test differences: Students who focus their attention on the important parts of the long oral argument transcript are likely to learn from these important parts better than those students who make notes about less relevant aspects of the text.

A possible explanation for the difference in the quality of student's note taking can be found in the use of the LARGO advice functions. Some of LARGO's feedback messages explicitly ask the students to consider specific important parts of the transcript in their diagramming, so it is likely that students who use the advice frequently also create notes that are about these parts of the argument (and have a chance to learn from them). Indeed, students in study 1 used the advice considerably more often than students in study 2, and also within study 2 a higher advice usage correlated with better results on some important post-test questions.

Reasons for the low usage of the advice function in 2007 may be connected to motivational issues: The extent to which users engage with a system depends on their specific goals. In the first study, the users were paid volunteers. As such they were more motivated to explore the system, to exercise the key features such as the graphical relations, the links between diagram and transcript, and the on-demand advice, and to take their time. The population in study 2 consisted of unpaid conscripts who had to use the system as a part of their course. They were inclined to use the system in the most convenient manner possible and thus ignored the central novel features. In many ways they used the system as a note taking tool with movable text boxes. Yet, the success or failure of an ITS, and particularly of one that offers its important features on demand as LARGO does, is governed by the extent to which its main features are exercised by the population. In the first study, the Low-LSAT students chose to exercise the novel features and showed performance gains. In the second study, the students did not do so consistently. Thus, the LARGO group derived fewer benefits from the system and performed identically. To make students engage with the beneficial features outside of the lab, it seems necessary to better integrate the tool into the classroom. In our second study, the use of LARGO was coordinated with the class but not a core part of it. Thus, it was more a lab session with non-volunteers than a real classroom activity. While the students were required to attend, they were not directly graded on their performance. The payoff for them lay in the preparation that the LARGO activities gave them for their future work. They were not specifically motivated to produce "good" graphs, to test the system features, or to do well in the post-test (which may also explain some pre-post performance drops). Rather, they were motivated to efficiently extract beneficial information for their writing assignments. If we want the students to use the advanced and important on-demand system functions, future studies of the LARGO ITS (and probably this result is also valid for other ITSs) should pay more nuanced attention to the specific motivations of the students, especially in real classroom situations.

CONCLUSION AND OUTLOOK

This paper presented a formative evaluation of LARGO, an ITS for legal argumentation which is based on argument diagrams. This system deals with two kinds of ill-definedness. Firstly, there is usually no fixed rule for deciding about a "best" or "correct" argument, but every argument is subject to debate and interpretation (domain ill-definedness); and secondly there may well be competing and equally good diagram representations for a complex oral/textual argument (representational ill-definedness). LARGO deals with these two kinds of ill-definedness by allowing students to diagrammatically reconstruct complex real-world arguments instead of making their own arguments (based on the assumption that they will learn from analyzing expert-level argumentation strategies), tolerating multiple argument graphs, and by giving feedback on diagrams in the form of self explanation prompts (i.e., by avoiding hard error messages).

A first pilot study carried out with volunteers in a first-year law school course provided some support for the research hypothesis that a diagrammatic language, combined with feedback that points out weaknesses and opportunities for reflection in students' argument diagrams, helps students learn to apply a general model of hypothetical reasoning, as they study transcripts of arguments made by highly-skilled experts. For lower aptitude students, the use of LARGO's diagramming and feedback functions was more effective than traditional note taking techniques. Specifically, within this group, those who used LARGO learned better to analyze new argument transcripts in the same area of the law, even when they studied the new transcript without the use of LARGO. They also learned better to reason about how a hypothetical might relate to a proposed test, a key element of hypothetical reasoning.

This was different in a second study with LARGO in which we tested the ITS as a mandatory part of a first semester law school class. Here, our results showed no evidence that the LARGO condition was better than the Control condition. The post-test was well-aligned with the instruction and we had sufficient statistical power. Our hypothesis was not confirmed, however. With the lack of overall differences between conditions, the study still showed some evidence that those students who engaged more with LARGO through the advice did better. This is consistent with our first study in which the paid volunteers used more of the LARGO features and benefited from them.

One lesson learned from the formative evaluation studies presented here which compare graphical representations in an intelligent tutoring system for the ill-defined domain of legal argumentation to a "traditional" text-based note taking tool is that the ITS may still be better, especially with lower-tier students, but that engagement and student motivation are decisive factors. This is especially true when "leaving the lab" and entering the classroom with ITS technology. It can make a big difference whether participation is voluntary or mandatory, and whether the students are also motivated to participate in a manner so that the key ITS features are used, especially if their usage is on-demand – which may be an adequate design choice for ill-defined domains. In our second study, the students' lack of motivation was clearly a problem – yet, a stronger incentive such as considering the result for the course grade was not feasible due to requirements of the Institutional Review Board who had to approve the study. Future studies will not only need to distinguish between the effects of external representations and intelligent feedback in order to more systematically investigate the factors of LARGO that do (and do not) help students, but also – on a way towards a regular classroom usage – will have to take these motivational factors into account.

This can probably be done through sanctions (grading the resulting graphs for credit), in-class support (e.g., lectures on the benefits of LARGO for the learning goals), or modifying LARGO in order to increase the student's engagement with the system even if their motivation to do so may be low. The current version of LARGO leaves many things to the users – the way they create the diagrams, how and if they link elements, and last but not least how often (if at all) they receive comments and feedback on their work. As previous research shows, this strategy may be problematic not only due to motivational aspects, but also because students frequently do not ask for help even though they could benefit from it (Aleven, Stahl, Schworm, Fischer, & Wallace, 2003). Yet, prompting the students with corrective feedback immediately after they make a mistake (as done by many successful ITS systems) is problematic in the ill-defined domain of legal argumentation. The LARGO feedback is given on demand on purpose, thus avoiding false error messages that are likely to occur in this ill-defined domain where it is often not clear whether a diagram correctly reflects an argument or not. False or inappropriate feedback would be very problematic if enforced on students also because the feedback LARGO gives is cognitively very demanding A reasonable alternative and a compromise between the two extremes, to be tested in further studies, could be to highlight diagram regions that LARGO could give feedback on – similar to the feedback in Andes (VanLehn, et al., 2005) – and thus make students more aware that help is available for a given issue, but not force them to take it or follow it immediately. We are currently also investigating other ways of increasing students' engagement with LARGO by letting them actually *make* arguments collaboratively in addition to *analyzing* them.

REFERENCES

Ainsworth, S. (1999). The functions of multiple representations. *Computers and Education*, 33(2), 131-152.

Aleven, V., Sewall, J., McLaren, B. M., & Koedinger, K. R. (2006). Rapid authoring of intelligent tutors for real-world and experimental use. In Kinshuk, R. Koper, P. Kommers, P. Kirschner, D. G. Sampson, & W. Didderen (Eds.) *Proceedings of the International Conference on Advanced Learning Technologies 2006* (pp. 847-851). Los Alamitos, CA: IEEE Computer Society.

Aleven, V., Stahl, E., Schworm, S., Fischer, F., & Wallace, R.M. (2003). Help Seeking in Interactive Learning Environments. *Review of Educational Research*, 73(2), 277-320.

Ashley, K.D. (2007). Interpretive reasoning with hypothetical cases. In *Proceedings of the 20th International Conference of the Florida AI Research Society* (pp. 387-392). Key West, FL: AAAI.

Ashley, K. D., Lynch, C., Pinkwart, N., & Aleven, V. (2008). A process model of legal argument with hypotheticals. In E. Francesconi, G. Sartor, & D. Tiscornia (Eds.) *Proceedings of the 21st International Conference on Legal Knowledge and Information Systems* (pp. 1-10). Amsterdam: IOS Press.

Carr, C. S. (2003). Using computer supported argument visualization to teach legal argumentation. In P. Kirschner, S. Buckingham Shum, & C. Carr (Eds.) *Visualizing Argumentation: Software tools for collaborative and educational sense-making* (pp. 75-96). London: Springer Verlag.

Chi, M. (2000). Self-explaining expository texts: The dual process of generating inferences and repairing mental models. In Glaser, R. (Ed.) *Advances in Instructional Psychology* (pp. 161-238). Mahway, NJ: Lawrence Erlbaum Associates.

Corbett, A.T., Koedinger, K.R., & Hadley, W. S. (2002). Cognitive Tutors: From the research classroom to all classrooms. In P. Goodman (Ed.) *Technology enhanced learning: Opportunities for change* (pp. 235-263). Mahway, NJ: Lawrence Erlbaum Associates.

Easterday, M., Aleven, V., & Scheines, R. (2007). 'Tis better to construct than to receive? The effects of diagramming tools on causal reasoning. In R. Luckin, K. Koedinger, & J. Greer (Eds.), *Proceedings of the 13th International Conference on Artificial Intelligence in Education* (pp. 93-100). Amsterdam: IOS Press.

Harrell, M. (2004). The improvement of critical thinking skills. In *What Philosophy Is*. Technical Report CMU-PHIL-158, Carnegie Mellon University, Department of Philosophy.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11, 65-99.

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

Lynch, C., Ashley, K., Pinkwart, N., & Aleven, V. (2007). Argument diagramming as focusing device: does it scaffold reading? In V. Aleven, K. Ashley, C. Lynch, & N. Pinkwart (Eds.) *Proceedings of the Workshop on AIED Applications for Ill-Defined Domains at the 13th International Conference on Artificial Intelligence in Education* (pp. 51-60). Los Angeles, CA.

Lynch, C., Ashley, K. D., Pinkwart, N., & Aleven, V. (2008). Argument graph classification via Genetic Programming and C4.5. In R. Baker, T. Barnes, & J. Beck (Eds.) *Proceedings of the 1st International Conference on Educational Data Mining* (pp. 67-76). Montreal. Published online at http://www.educationaldatamining.org/EDM2008/index.php?page=proceedings

Mitrovic, A., Martin, B., & Suraweera, P. (2007). Intelligent tutors for all: The constraint-based approach, *IEEE Intelligent Systems*, 22(4), 38-45.

Paolucci, M., Suthers, D., & Weiner, A. (1996). Automated advice-giving strategies for scientific inquiry. In C. Frasson, G. Gauthier, & A. Lesgold (Eds.) *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 372-381). Berlin: Springer Verlag.

Pinkwart, N., Aleven, V., Ashley, K., & Lynch, C. (2006a). Toward legal argument instruction with graph grammars and collaborative filtering techniques. In M. Ikeda, K. Ashley, & T. W. Chan (Eds.) *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 227-236). Berlin: Springer Verlag.

Pinkwart, N., Aleven, V., Ashley, K., & Lynch, C. (2006b). Using collaborative filtering in an intelligent tutoring system for legal argumentation. In S. Weibelzahl, & A. Cristea (Eds.) *Proceedings of Workshops held at the 4th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems. Lecture Notes in Learning and Teaching* (pp. 542-551). Dublin: National College of Ireland.

Pinkwart, N., Ashley, K., Lynch, C., & Aleven, V. (2008). Graph grammars: An ITS technology for diagram representations. In H. Chad Lane, & D. Wilson (Eds.) *Proceedings of the 21st International Conference of the Florida AI Research Society* (p. 433-438). Coconut Grove, FL: AAAI.

Reed, C., Walton, D., & Macagno, F. (2007). Argument Diagramming in Logic, Law and Artificial Intelligence. *The Knowledge Engineering Review*, 22, 87-109.

Reese, L., & Cotter, R. (1994). *A Compendium of LSAT and LSAC-Sponsored Item Types 1948-1994.* Law School Admission Council Research Rept. 94-01. http://www.lsacnet.org/Research/Compendium-of-LSAT-and-LSAC-Sponsored-Item-Types.pdf

Schank, P, & Ranney, M. (1995). Improved reasoning with Convince Me. In I. R. Katz, R. L. Mack, L. Marks, M. B. Rosson, & J. Nielsen (Eds.) *Proceedings of the International Conference on Human Factors in Computing Systems* (pp. 276-277). New York, NY: ACM.

Schworm, S., & Renkl, A. (2002). Learning by solved example problems: Instructional explanations reduce self explanation activity. In *Proceedings of the 24th Annual Conference of the CogSci Society* (pp. 816-821). Mahwah, NJ: Lawrence Erlbaum Associates.

VanLehn, K., Lynch, C., Schulze, K., Shapiro, J.A., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., & Wintersgill, M. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence and Education*, 15 (3), 147-204.

Voss, J. (2006). Toulmin's model and the solving of ill-structured problems. In D. Hitchcock & B. Verheij (Eds.) *Arguing on the Toulmin Model: New Essays in Argument Analysis and Evaluation* (pp. 303-311). Berlin: Springer Verlag.

Voss, J. F., & Means, M. L. (1991). Learning to reason via instruction in argumentation. *Learning and Instruction*, 1, 337-350.