Argument diagramming as focusing device: does it scaffold reading?

Collin Lynch[1], Kevin Ashley[2], Niels Pinkwart[3] and Vincent Aleven[4]

[1]*Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA 15260, USA (collinl@cs.pitt.edu)*

[2]*Intelligent Systems Program and School of Law, University of Pittsburgh, Pittsburgh, PA, 15260 USA (ashley@pitt.edu)*

[3]*Clausthal University of Technology, Julius-Albert-Str. 4, 38678 Clausthal-Zellerfeld, Germany (niels.pinkwart@tu-clausthal.de)*

[4]*Carnegie Mellon University, HCI Institute, 5000 Forbes Avenue, Pittsburgh PA 15213, USA (aleven@cs.cmu.edu)*

Abstract: In this paper we report on a study of attention and student recall in our ITS LARGO. The system was employed in a study of graphical markup in legal education. Students in the study were divided into two groups, one employing the graphical tutoring environment, and the other traditional text notes and highlighting. Post-test comparisons between the two showed gains among the incoming students who had scored lower on a standardized Law School Admissions Test (LSAT). We argue that the system and its graphical prompts were effective in guiding the students to the relevant textual portions and that they showed some gains in focus of attention.

Keywords: Ill-defined domains, note-taking, attention, self-explanation.

Introduction

An ill-defined problem-solving task is one in which (1) the problem does not have a definitive answer, (2) the way in which the problem-solver solves the problem depends on how he conceptualizes it, and (3) problem-solving involves identifying relevant concepts and mapping them onto the situation to be solved [10]. Deciding how to resolve a legal dispute is an ill-defined task. Reasoning with hypotheticals is a strategy for dealing with that. Each participant (i.e., the contending advocates, the deciding judges) may propose a different, perhaps inconsistent but often reasonable solution. The alternatives often evidence differences in the ways in which the participants conceptualized the problem or applied those legal concepts to the problem's facts. In applying the concepts, legal reasoners often draw analogies between the problem's facts and past or hypothetical cases; these analogies map legal concepts that apply in the hypotheticals or precedents onto the present case's facts to help draw and justify conclusions.

This work focuses on legal problem-solving at the Supreme Court of the United States (SCOTUS). A feature of problem-solving at this level are oral arguments before the Court. Each side in a case gets thirty minutes to address the issues before the Court; the arguments are recorded and later published. In it an advocate for one side proposes a rule or test for deciding the case in favor of his client. Justices in turn pose hypotheticals in order to probe the proposed rule. The hypotheticals help the Justices to understand what the proposed test means, whether it is consistent with past decisions, and how well it implements and reconciles the conflicting legal policies and principles. Legal reasoning with hypotheticals is one of the tools Justices have for mapping legal concepts from past decisions and applicable statutory and constitutional provisions onto the problem's facts and adjusting the mappings to account for underlying legal policies and principles.

As such SCOTUS oral arguments provide good examples of reasoning with hypotheticals for law students to study. Law students are exposed to, and sometimes participate in, Socratic dialogues in classes from which they should learn to reason about legal rules with cases and hypotheticals. The SCOTUS oral arguments are potentially a pedagogically valuable source of examples of this kind of reasoning. They are realistically complex, often highly dramatic, and they are written down which

facilitates studying them at some length. On the other hand, they are an underutilized pedagogical resource. Law professors may employ SCOTUS oral arguments to teach lessons about the substantive law of an area, but they do not generally use them as examples of argumentation methods. While traditional legal education encourages students to make and respond to arguments, it does not provide much explicit support for reflecting on the process.

The LARGO program attempts to redress that failing by helping law students reflect on SCOTUS oral arguments as examples of legal argumentation. An intelligent tutoring system (ITS), it teaches legal reasoning with hypotheticals by helping students to represent selected elements of these examples of expert legal arguments in diagrams (Other legal ITSs include CATO and CATO-Dial [2,6].) The elements include an advocate's proposed test for deciding a legal case, Justices' hypothetical examples posed to probe the test, and the advocate's responses to the hypotheticals. Students identify these elements in the text, represent them in a diagram, providing their own reformulations of the text, and link the elements graphically indicating certain dialectical relationships among them [12]. Given the ill-definedness of the task, and the subjectivity of interpreting the textual argument, LARGO cannot simply teach by identifying "right" and "wrong" answers. Instead it provides feedback based upon expert markup and an understanding of common dialectical patterns. This hint mechanism will be described below.

The value of note-taking has long been recognized in legal education [9] but the focus has always been on text notes. Graphical notes have been shown to be beneficial through their ability to focus the student's attention on relevant portions of the text [13]. A similar effect has been noted for ITS feedback [1]. Graphical argument representations have been studied in philosophy [15] and legal education [7]. Unfortunately the results were inconclusive.

In an experiment, we compared the LARGO program with a more traditional text-highlighting-and-note-taking word-processing environment that focused students on the same elements and relationships of hypothetical legal reasoning but without the diagramming or feedback. We found evidence that students with lower LSAT scores benefited the most from LARGO and its support of graphically diagramming arguments [12]. These students using LARGO learned some targeted skills of hypothetical legal reasoning better than comparable students in the control group (the Text-only group).

We have begun to attempt to explain why LARGO has benefited such students in the Diagram group. This paper reports the results for our initial hypotheses in explaining the data, that (1) students in the Diagram group, with LARGO's support, are more successful in finding and attending to pedagogically-relevant portions of the text than students in the Text-only group. In particular, (2) students in the Diagram group with lower LSAT scores, which may indicate lower reading skills, benefit more from LARGO's support in finding and attending to important portions of the text than higher LSAT students.

In the next section, we describe LARGO's instruction about hypothetical reasoning and illustrate it with an example of a student's diagram of excerpts of an oral argument. Following that we describe the former experiment and illustrate the output of a student in the Text-only group. In the Empirical Evaluation of Attention section we describe our current empirical evaluation comparing the portions of the text students attended to in the Text group vs. the Diagram group, including the way we operationalized that comparison. In the subsequent sections, we present the results, discussion, and our conclusions.

LARGO Instruction.

In our study, law students read oral argument transcripts from SCOTUS. Figure 1 contains an example of the tests and hypotheticals encountered in such arguments drawn from the oral argument in *Burnham v. Superior Court of California*, 495 U.S. 604 (1990). The left column contains the text of the argument with line numbers. Mr. Sherman makes arguments on behalf the "petitioner" in the case, Dennis Burnham; "QUESTION:" indicates a Justice's question.

Here are the facts of the case. After Burnham and his wife decided to separate, she moved to California with their two children. In January, 1988, Mrs. Burnham filed suit in California for divorce. Later that month, Burnham visited California on business and to visit his children. Upon returning one of them to Mrs. Burnham's home, he was served with her divorce petition. Later that year, he appeared in California Superior Court to assert that the courts there lacked *personal jurisdiction* over him. Personal jurisdiction, a technical legal concept first year law students encounter in

their "Legal Process" course, means a court's power to require a person or corporation to appear in court and defend against a lawsuit. Burnham argued that his contacts with California, consisting only of a few short visits to conduct business and visit his children, were insufficient to grant the courts there jurisdiction of his person under the Due Process Clause of the Fourteenth Amendment, which guarantees certain minimum procedural safeguards against the arbitrary exercise of government power. Conflicting with that principle is the principle that a state may redress wrongs committed within or affecting residents of the state. The California Superior Court denied his motion, and the SCOTUS agreed to review that decision. The Court affirmed the lower court decision, but could not agree on a majority opinion.

Oral argument excerpts	**Argument Move According to Model of Hypothetical Reasoning**
5. The issue presented here is whether a state can exercise personal jurisdiction over a nonresident defendant who was personally served while present in the state if that defendant does not otherwise have sufficient contacts with the state to satisfy the minimum contacts test announced in International Shoe. 11. We're here today to ask you to instruct the courts of this land otherwise, to give effect to what the Court said in Shaffer, that personal jurisdiction in all cases must be tested by the minimum contacts test.	**→ Proposed test of Mr. Sherman for Petitioner Burnham**
15, 17. QUESTION: Mr. Sherman, even if you are correct that some minimum contact is necessary for personal jurisdiction, wouldn't the transitory presence within the state of someone meet that test -- in a good many instances?	**← J.'s hypo**
18. MR. SHERMAN: I think not, Your Honor. And it's important to distinguish --	**→ Response: distinguish cfs/hypo**
19, 21. QUESTION: I would have thought so and that perhaps someone who voluntarily enters a state to transact some business or to visit there might well meet whatever minimum contacts are -- required.	**← J.'s hypo**
22, 24. MR. SHERMAN: On that -- on those facts, yes. If he were just passing through momentarily, say, stopping over on his way to Hawaii, not conducting any business or -- classically flying over --	**→ Response: distinguish cfs/hypo**
25, 27, 29. QUESTION: You have a different situation if someone is flying over the state, overhead -- and is served in mid-air than you do with someone in your client's --- position.	**← J.'s hypo**
30. MR. SHERMAN: But the question that your hypothetical poses is what kinds of contacts would be sufficient under the minimum contacts test for somebody who was not in the state very long. And the answer to that would depend upon applying the minimum contacts test and typically the cause of action has to be related to or rise out of contacts that the defendant has.	**→ Response: distinguish cfs/hypo; modify test to exclude hypo.**

Figure 1: Oral Argument Excerpts and Argument Moves from Burnham.

The right column of Figure 1 lists the argument moves in our model that correspond to the assertions. For more on the model see [3,4,5]. Mr. Sherman proposes a test deciding the issue in favor of Mr. Burnham, namely there is no personal jurisdiction without a showing of adequate minimum contacts. The Justices challenge that test with hypotheticals, asserting that there *are* adequate minimum contacts in this case. Mr. Sherman distinguishes the hypotheticals and finally asserts a meaningful distinction with a corresponding test modification. Even if minimum contacts exist, he argues, the cause of action (the subject of the suit) must arise from them to have personal jurisdiction.

Figure 2 shows an example of the LARGO interface. The argument transcript is on the top left. On the right is a workspace for creating the diagram using the palette of representation elements at the bottom left. Students create graphs representing an argument exchange in the transcript by dragging the elements from the palette to the workspace. Elements exist for representing the current fact situation, proposed tests (and modifications), hypotheticals, and various relations among them (e.g., modifying a test, distinguishing or analogizing a hypothetical, and a general relation). Students also link the elements in their diagrams to passages in the transcript via a highlighting feature.

The diagram in the workspace at the right of Figure 2 is a student's representation of some of the excerpts of the *Burnham* argument (see Figure 1.) The student has represented the Justice's "served-while-flying-over" hypothetical and linked it via the highlight function to a portion of the transcript including lines 25, 27, and 29. The student has also identified a version of Mr. Sherman's test (top) which gets modified to a version that roughly accounts for the additional contact-centric limitation imposed in line 30.

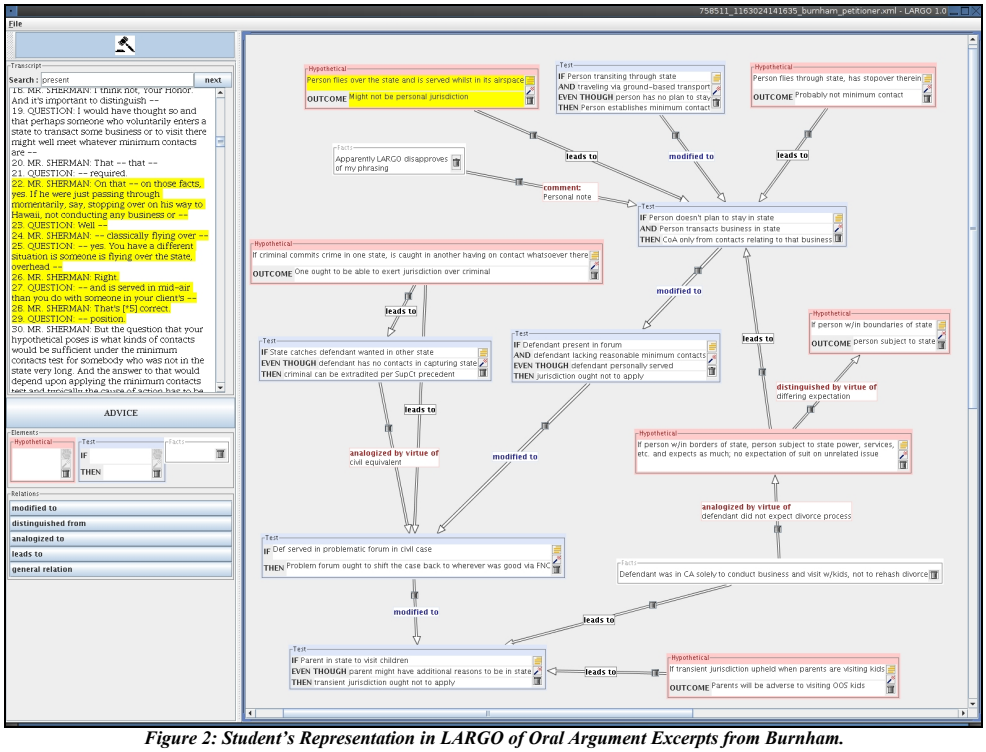

Figure 2: Student's Representation in LARGO of Oral Argument Excerpts from Burnham.

Students may obtain feedback on their developing diagram by clicking on the Advice button. This brings up a palette with up to three hint choices with titles such as "Reflect on the role of hypotheticals in the transcript". Clicking on a title brings up a more detailed message. All of LARGO's help is provided by request only.

The feedback leads students to review the text of the oral argument in at least four ways: (1) Some feedback identifies regions in the text where the student's graph lacks elements corresponding to those in an expert's markup of the argument transcript. Prior to a transcript's use in LARGO, an expert marks passages of interest such as those shown in Figure 1. This type of hint points the students to a larger region than the actual element and informs them that an item of interest is present in it. (2) Other feedback identifies parts of the diagram where the relations among elements do not correspond to the "standard" model. For instance, hypotheticals are commonly analogized to or distinguished from one another and the current fact situation A student's graph may fail to show such relationships or may indicate uncommon relationships (e.g., analogizing or distinguishing a test and a hypothetical.) This type of feedback may lead students to reexamine the portions of the text that embody the elements and their relations. (3) Some system advice asks students to compare their test formulations to examples from other students or the professor; this may lead students to reexamine the text containing an advocate's test as they consider which conditions to include and how abstractly to characterize them. (4) Finally, some LARGO advice identifies a standard dialectical pattern or node configuration in the students' model and asks them to reflect on how the pattern bears on the argument's merits and whether a different decision might have been more appropriate. The student may be pointed to a proposed test that led to a hypothetical which in turn prompted the advocate to modify the test (e.g. top of figure 2). Such self-explanation prompts [8] lead them to reread the initial text and, on occasion, to modify their representations. This has been shown to be effective when studying examples [14].

Fall 2006 Experiment.

In fall 2006, we conducted an experiment to investigate to what extent LARGO can lead to better learning than a traditional purely text-based alternative. The alternative tool simulates the

"traditional" process of examining the argument transcript with a notepad alone by allowing students to highlight selected portions of the transcript text and enter their notes in a text pane. Figure 3 shows a screenshot of the tool as it has been used by one of the study participants to annotate the transcript of the oral argument in Burnham.

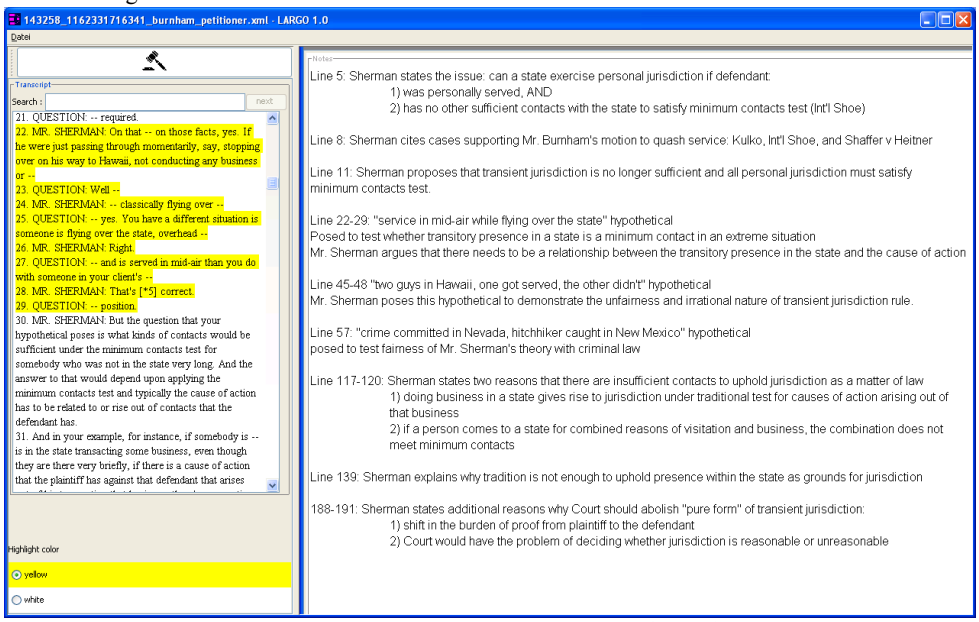

Figure 3. Screenshot of control condition (text tool).

The experiment was conducted with first-year students at the University of Pittsburgh's School of Law. All were volunteers and were paid $80 on completion. All cases examined in the study centered on questions of personal jurisdiction and were part of their coursework. The students were assigned randomly to the conditions. 38 students began the study, 28 completed.

The experiment consisted of four 2 hour sessions over a single one month. The first involved a pre-test and an introduction, with example, to the dialogue model and the appropriate system. In session 2 students used the systems to examine extracts of the Burnham oral arguments. In Session 3 they repeated the process with *Burger King Corp v. Rudzewicz* 471 U.S. 462 (1985). Experimental students used LARGO for note taking while the control subjects employed the text system. Both groups were instructed to take notes on the argument using the tool.

Our main hypothesis was that LARGO's graphical and advice tools would help students better identify and mark up the argument components, leading to better learning of argument skills. A first analysis of the results has been published in [12]. On average, the experimental group did better in the post-test than the Control group, yet the difference was not statistically significant ($t(1,26)=.92$, $p>0.1$). Dividing the students into three groups according to their Law School Admission Test (LSAT) scores, revealed that the "Low" experimental students benefited most from LARGO. This group scored significantly higher than their Control counterparts in several categories of post-test items (although not overall) including argumentation about a near-transfer problem, questions on a novel personal-juristiction case and questions asking them to evaluate argument components [12]. The results support our hypothesis (though they perhaps fall short of decisive confirmation). For students who do not (yet) have the ability to learn argumentation skills based on reading alone, using LARGO led to significantly better learning of argumentation skills than the traditional note-taking techniques. For the more advanced/skilled students, LARGO was neither better nor worse than traditional methods.

Empirical Evaluation of Attention.

Our post-test results, while intriguing, do not provide a definite measure of LARGO's utility. We thus undertook a more detailed attention analysis. This analysis was performed on the data files and logs

produced by the students during the study in order to understand what aspects of LARGO (visual representation, linking of graph to transcript, or advice) led to the observed difference between conditions. We investigated whether students using LARGO were more successful in finding and attending to the important portions of the text (i.e., relevant test and hypothetical formulations in the argument) than students who use the text-tool. If there is a difference between the conditions in terms of finding and locating relevant textual items it could partially explain the post-test differences: Students who, assisted by the tool, focus their attention on important parts of the long oral argument transcript, are likely to learn the important parts better.

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. We did not consider their facility at identifying the relevant legal issues or relationships between these elements. As we noted above the students were assigned to examine three cases during the study. Their examination of the intro case (California v. Carney, 105 S. Ct. 2066 (1985)) was guided by a document which presented the domain model as well as a step-by-step sequence of appropriate analyses. The remaining two cases, *Burnham* and *Burger King* were analyzed without such guidance. And we focus on them here.

In preparation for this study an expert legal instructor marked up the redacted transcripts of each argument. This individual identified a set of important Tests and Hypotheticals in each oral argument. This markup did not include defining an ideal graph or summary statements of each element, only the identification of relevant regions. This resulted in a list of 33 regions over the two cases. Sixteen of these regions (*Core* set) were encoded into Largo for use in providing hints of type 1 (see above). The remaining 17 were reserved for this study and designated as the *Test* set.

Our goal was to provide both a basis for hinting (Core), and a baseline (Test) for comparison. For the Graph condition these two sets form a test-train split of a type commonly employed in Machine Learning (ML) [11]. As noted in the introduction this process of markup is an ill-defined one. We hold no expectations that the students overall analysis will match ours. However we feel that it is neither illogical nor extreme to expect them to locate the same statements of tests and hypotheticals as a legal expert, and we expect good students to perform better. While there are complicating factors, we argue that this is an appropriate methodology and one that draws on the relevant literature.

(Entries) In order to effectively compare the two groups we defined a standard baseline unit of student work. We therefore defined the *Note* as a single atomic reference or notation made by the students. For LARGO students a note is a single graph node or relation. The test and hypothetical nodes may be linked to the graph. A node is *location-relevant* if it is linked to one of the Core or Test locations irrespective of type. It is *type-relevant* if it links to the location and is of the correct type.

The graph shown in figure 2 contains 29 notes. Including relational links and fact nodes in the class of `notes' penalizes the graph subjects for making those entries as they cannot increase the success measures only decrease them. We opted to include them for three reasons: (1) the students' task was to markup 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 graph condition would bias the results in their favor as no viable standard was available for discounting text notes in the same way.

A Text note is defined as a single paragraph entry that may be accompanied by a highlight. Such a note is *location-relevant* if the text explicitly references some key transcript portion by line number or via a highlight. It is *type-relevant* if it explicitly identifies the type of the location in text. Figure 3 contains 9 textual notes. The 4th, 5th, and 6th all specify a type. The structure of these notes is similar to those given as examples in Session 1. We defined note in this way to ensure that the text and graph subjects employed roughly the same amount of cognitive effort when making each note.

(Measurement) We will focus the remainder of our discussion on three Data measurements (Time, Help & Work), and three Success measures (Efficiency, Precision, & Recall) commonly employed in ML applications. We will discuss all the measurements on a case basis (e.g. Time spent on Burnham) and overall. *Time*, is a measure of the time spent on task. *Help* reflects the number of advice requests the student made and the number of times they followed up with specific advice (applicable only to LARGO students). *Work* is a measure of the number of notes made by each student. For the LARGO students we counted each node and edge. For the text students we approximated the total number of notes by using the number of highlights or text notes whichever was largest. Thus we kept the count linked to distinct note-taking acts. While this may undercount

slightly we think that it is a viable choice.

The 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 that were located by the student out of the total number. *Efficiency* is the rate at which the students located the relevant notes. *Precision* is the number of relevant elements located versus the amount of work done. These definitions vary somewhat from those typically used in ML but we find them more appropriate here. We calculated each measurement with respect to the Core and Test sets and overall. We present the results below.

$$\frac{(Key\ Spots\ Found)}{(Total\ Key\ Spots)} \qquad \frac{(Key\ Spots\ Found)}{(Total\ Notes\ Made)} \qquad \frac{(Key\ Spots\ Found)}{(Time\ on\ Task)}$$

Recall Precision Efficiency

Figure 4: Success Measures.

(Hypotheses) Extant research on the benefits of graphical notes asserts that the students using them will be better able to 'focus in' on the relevant material. As such we have the following hypotheses: *(h0)*: students in the graph condition will have higher *recall* than their textual counterparts. *(h1)*: students in the graph condition will be more *efficient.* And *(h2)* students in the graph condition will have higher *precision* than their text counterparts. We discuss data relating to these hypotheses below.

Results

During the study we controlled for time on task and there was no significant difference between the two conditions either in terms of total study time or time spent on each case. The only variation occurred within the High student pool. There the Text students spent significantly more time overall ($t(5.09)=33.67$ $p < .0.0069$[1]) and on a per-case basis ($t(13.2)=3.71$ $p < 0.002$ for Burnham and Burger King). We discuss to the significance of these differences below.

There was no overall difference between the conditions in terms of the work done. There was, however, a case-specific difference. On Burnham there was a trend ($t(12.3)=1.75$ $p < 0.05$) in favor of the Text condition indicating that they did more work. This same pattern appeared in Burger King but was very significant ($t(12.43)=2.7$ p 0.008). Unlike Time there was a within-condition trend with the graph condition doing more work on Burger King than on Burnham ($t(27.87)=-1.7$ $p < 0.05$). As before the High Text students did more overall ($t(5)=3.71$ $p < 0.01$) and on a case basis ($t(5)=4.33$ $p < 0.01$ and $t(5)=3.9$ $p < 0.01$ for both cases).

Table 1 shows an overall comparison between the text and graph groups. For this and other tables p-values are not shown. Regular text indicates $p < 0.025$ while italics indicates $0.025 < p < 0.1$. As you can see in Table 1 the Graph condition outperformed the Text condition in terms of Efficiency, Precision, and Recall on the *Core* set both in terms of the location and type-relevance standards. They were also were more precise at locating the relevant elements and showed greater efficiency, precision and recall about typing them. This pattern continued on the full set. Interestingly, the Text condition performed better at locating the relevant entries but not at assigning types to them on the *Test* set. This pattern was observable on Burnham alone (Table 2) but not for Burger King. Here the graph condition was dominant and, while not outperforming the Text condition in terms of Efficiency and Recall, did outperform them in terms of Precision (Table 3).

[Table 1]	Core	Test	All
Eff Located	T<G t(25.8)=-5.3[2]	T>G t(25.3)=3.34	
Typed	T<G t(25.17)=-7.6		T<G t(23.7)=-5.56
Prec Located	T<G t(25.3)=-4.3	T>G t(26)=2.5	*T<G t(26)=-1.8*
Typed	T<G t(25.4)=-6.7		T<G t(26)=-4.9
Rec Located	T<G t(12)=-6.8	T>G t(24.4)=3	
Typed	T<G t(12)=-11.1		T<G t(17.3)=-7.3

Table 1: Overall Condition Comparison.

1 Unless otherwise stated all test values are from Welch's Two-sample 1-sided t-test.

2 T<G means that the text students scored below the graph students.

	[Table 2] Burnham.			[Table3] Burger King.		
	Core	Test	All	Core	Test	All
Eff L.	T<G t(25.6)=-4.7	T>G t(20.5)=4.4		T<G t(26)=-5.1		*T<G t(25.3)=-1.8*
T.	T<G t(23.5)=-6.7		T<G t(19)=-5.2	T<G t(24.3)=-7.5		T<G t(25.8)=-4.8
Prec L.	T<G t(18.5)=-5.9	T>G t(25.2)=2.3	T<G t(23.7)=-5	T<G t(17.6)=-7.6	T<G t(18.6)=-2.4	T<G t(18)=-5.1
T.	T<G t(18)=-7.1	*T<G t(22)=-1.6*	T<G t(21.5)=-7.8	T<G t(16.8)=-9.1	T<G t(20.8)=-3.2	T<G t(18)=-6.4
Rec L.	T<G t(12)=-5.9	T>G t(20.5)=4		T<G t(12)=-5.7		T<G t(22.3)=-2.2
T.	T<G t(12)=-9.5		T<G t(15.6)=-6.3	T<G t(12)=-9.8		T<G t(21.1)=-6

Tables 2 & 3: Case by case comparison of condition.

Further analysis of within-condition variations revealed that neither group performed better overall on either case. While both groups were more efficient and precise on Burnham than Burger King on the Core set, this was not reliably the case for the Test set or the full sets. Interestingly both groups performed better on the later case with respect to the Test set (Tables 4 and 5).

When analyzing our study results we split the students into Low, Medium, and High groups based upon their LSAT scores [12]. In that analysis, the Low LSAT Graph students gained more than their Low Text counterparts while the Medium and High students showed no across the board distinctions. We further analyzed the overall variation between the groups with respect to the five measures. The Low groups showed no significant difference in terms of time-on-task and amount of work done while the High Text group both took significantly more time than their Graph counterparts overall ($p < 0.0069$) and on a case basis (Burnham $p < 0.001$, Burger King $p < 0.01$). This was also true for work (Overall $p < 0.01$; Burnham p0.001; Burger King $p < 0.006$).

	[Table 4] Text.			[Table 5] Graph.		
	Core	Test	All	Core	Test	All
Eff L.	B>G t(21.4)=2.1[3]	B<G t(22.8)=-2.2		B>G t(26.4)=3.2	B<G t(20.8)=-4.5	
T.		B < G t(22.2)=-3.7		B>G t(26.4)=3.2	B<G t(21.6)=-4.7	
Prec L.	B > G t(16.9)=3.5		B>G t(20.9)=2.1	B>G t(21)=3.4	B<G t(20.5)=-3.7	
T.	B > G t(16.3)=2.7	B<G t(18.7)=-2.4		B>G t(21)=3.4	B<G t(20.5)=-3.7	
Rec L.						
T.						

Tables 4 & 5: Between case comparison for the Text and Graph Conditions.

Comparisons between the Low groups on the three success measures (Table 6) closely parallels the overall breakdown between the groups. The High students showed more consistent variation (Table 7) in favor of the Graph condition with the graph students outperforming their text counterparts across the board on the Core set and having higher type performance on all three measures. They did not, however, display the same variation on the Test set. There the only variation was the text students' higher recall of locations but not types.

	[Table 6] Low LSAT.			[Table 7] High LSAT.		
	Core	Test	All	Core	Test	All
Eff. Fnd.	T<G t(7.8)=-3.4	T>G t(7)=2.6		T<G t(5.3)=-9.7		T<G t(5.8)=-3.3
Typ.	T<G t(7.6)=-6.6		T<G t(7.5)=-4.9	T<G t(5.2)=-11.7		T<G t(4.6)=-8.1
Prec. Fnd.	*T<G t(7.6)=-1.6*	*T>G t(7.9)=2*		*T<G t(5)=-17.6*		T<G t(2.24)=-5.4
Typ.	T<G t(6.7)=-3.5		T<G t(7.7)=-2.6	T<G t(5)=-16.3		T<G t(4)=-7.6
Rec. Fnd.	T<G t(4)=-3.6			T<G t(5)=-7.3	*T>G t(2.1)=2.9*	
Typ.	T<G t(4)=-9.8		T<G t(5.6)=-5.3	T<G t(5)=-9.8		T<G t(2.4)=-6.8

Tables 6 & 7: Cross-condition comp for the Low and High LSAT students.

3 For this and Table 4 B<G means that the measure was higher for Burger King than Burnham.

Analysis of the help usage revealed no significant variation in help usage from case to case. Nor was there any significant difference in help usage between either the Low or High students. Indeed the only notable variation detected was in the amount of help selection between the High group and the remaining students (p 0.03). That is, the High Graph students clicked on the help button as often as their peers but followed up on that by selecting one of the choices less often.

Discussion

The lack of clear overall differences between the conditions in terms of time on task indicates that the graphical tools imposed no additional cognitive load. Despite every law student's unfamiliarity with graphical representations they took no more time to utilize the graphical tools. If the tools were overwhelmingly complicated, we would expect some students, especially in the low group, to perform worse and this was not the case. Similarly the equality of work performed suggests that the students were, for better or worse, expending the same amount of effort in either condition. Thus any gains attributable to the system are due not to load reductions but better use of time and effort.

This hypothesis is supported by the success measures. Our original hypotheses that the LARGO condition would dominate in the success measures held in part. When measured both overall and case-by-case, the LARGO condition was clearly dominant on the Core set. This was true both in terms of location-relevance and the higher type-relevance standard. This suggests that the advice was effective though it did not explicitly state the missing tests or hypotheticals only a region of interest.

This dominance did not hold when measuring against the Test set, consisting of elements that LARGO did not point students to. There the Text students were dominant with respect to location-relevance and the two conditions were equal in terms of type-relevance. This was true both overall and for Burnham save for the LARGO condition's increased type precision. This reversal was not present on Burger King where the two groups were equal in terms of Efficiency and Recall and the LARGO condition was dominant in terms of Precision.

In our opinion this can be explained by three related factors. Firstly we believe that the students within the graph condition may initially have engaged in some form of help-dependence and relied overmuch on the system to point out all essential components on Burnham and less so on Burger King. This would explain their clear success on the Core and mixed success on the Test sets. Further analysis will be necessary to confirm this.

Secondly the distinction between the Core and Test sets was not random as is the case in most ML applications. The Core elements were arguably more important to the dialogue than those of the Test set. It is possible that the students, making the same subjective assessment, focused more effort on the Core components. Moreover, the two sets were unequally distributed in terms of Tests and Hypotheticals with the bulk of the tests located in the Core set. Thus measurement of overall improvement was confounded somewhat with set measurement.

Thirdly, in examining the text students' notes we observed that a high proportion of their focus was on "action" or "concept" notes rather than relations. Thus while they were informed of the value of distinctions and other relationships they took few notes about them and focused instead on identifying relevant tests, hypotheticals and legal concepts. While we have not yet fully coded the notes we believe that the text students spent more time "making dots" rather than connecting them. Thus they have a higher proportion of candidate tests and hypotheticals to actual tests and hypos than their LARGO counterparts. This gave them an initial boost when it came to location-relevance but not type-relevance. By Burger King this gap had been removed or even reversed.

This hypothesis is somewhat complicated by the between case comparison for the conditions. Both conditions performed equally or better on Burnham with respect to the Core set and equally or better on Burger King with respect to the test set. Clearly in both cases the students were gaining familiarity with the model and were more willing to move beyond the most "central" elements. While this might be taken to suggest that the conditions learned equally, the lack of improvement in location-relevant precision by the text students as compared to the graph students and the overall dominance of the LARGO students on Burger King suggests otherwise. More data is required.

We believe that the results are consistent with our low-LSAT versus high-LSAT post-test results. While the low students clearly followed the same overall pattern of the group, the high students did not. There the LARGO condition was dominant on the Core and full sets and equal (save for location-relevant recall) on the Test set. This in spite of the fact that in this case the high text students both performed more work and spent more time than their graph counterparts. We believe that this demonstrates effective use of the system by the high LARGO students and a wiser

recognition of where to focus their efforts. In our opinion the post-test results suffered from somewhat of a ceiling effect thus washing out any apparent variation save between the low students. We further note that our "low" LSAT scores are in fact in the middle or upper middle segment of the average law school population. As such they are not representative of what truly "low LSAT" students might do. At present we are planning to conduct such a test this summer.

Conclusions.

The results that we noted above are positive and support our position that LARGO is beneficial for the students. The gains observed in the study can be partially explained by the findings presented in this paper – by means of reflective prompts, LARGO is able to help the students focus their attention on the important elements of the argument and helps them better to recognize tests and hypotheticals. Since thinking in terms of tests and hypotheticals is central to the argument model, this would mean that they had an important foundation in place on which to build further understanding of the model. Far from requiring a great deal of initial ramp-up, the students were able to adapt to the tools fairly quickly and showed improvements contrary to some expectations. While some potential system improvements were suggested by this study the system was largely successful.

As we noted above there were several complicating factors in this study that we expect to address in future studies. We will be conducting a study with genuinely 'low LSAT' students this summer. We plan to retool our post-test, provide more study cases, and to strive for a more appropriate distinction between the test and train sets. The last is of course the most difficult to address. Unlike individual utterances in a text-to-speech scenario, the tests and hypotheticals in our cases are not entirely independent. They often refer to one another, are modified, reappear and so on. Some may genuinely be considered more or less important than others although debate rages over which. Additionally some cases favor a large proportion of hypotheticals to tests, and some only a few or even none. Thus we will not be able to achieve a truly random split.

The fact that this method works in an ill-defined domain is intriguing. In such domains the challenge for an ITS is often to balance between providing too much structure and too little. While researchers agree that this problem exists there is little agreement on where the sweet spot may be found. One mechanism commonly used to address such a question is the fading of help from explicit to general hints, to no hints over time. Our system by contrast provides solely general information which proved useful to low, medium and high students without undue constraint. We plan further investigations along these lines as part of our subsequent work.

References

[1] Anderson, J.; Corbett, A.; Koedinger, K.; Pelletier, R.: "Cognitive tutors: Lessons learned." *The Journal of Learning Sciences* 4 (1995) 167-207.

[2] Aleven, V. (2003). Using Background Knowledge in Case-Based Legal Reasoning: A Computational Model and an Intelligent Learning Environment. Artificial Intelligence 150, 183-237.

[3] Ashley, K.D. (1990) *Modeling Legal Argument: Reasoning with Cases and Hypotheticals.* MIT Press. Cambridge.

[4] Ashley, K.D. (2006). "Hypothesis Formation and Testing in Legal Argument." Invited paper. *Inst. de Investig. Jurídicas 2d Intl Meet. on AI and Law*, UNAM, Mexico City. April

[5] Ashley, K.D. (2007). "Interpretive Reasoning with Hypothetical Cases." In *Proc. 20th Int'l FLAIRS Conference, Special Track on Case-Based Reasoning,* Key West, May.

[6] Ashley, K.D., Desai, R. and Levine, J.M. (2002) "Teaching Case-Based Argumentation Concepts using Dialectic Arguments vs. Didactic Explanations". In Proc., ITS '02 (S.A. Cerri, G. Gouardères, F. Paraguaçu, ed.) pp. 585-595. Springer: Berlin.

[7] Carr, C. (2003). "Using Computer Supported Argument Visualization to Teach Legal Argumentation." In *Visualizing Argumentation*, 75-96. London, Springer.

[8] Chi, M. (2000) "Self-explaining expository texts: The dual process of generating inferences and repairing mental models." In *Advances in Instructional Psychology.* Glaser R. (Ed.) Lawrence Erlbaum

[9] Llewellyn, K.N. (1960) *The Bramble Bush; on Our Law and its Study.* Oceana Publications. New York.

[10] Lynch, C., Ashley, K., Aleven, V., & Pinkwart, N. (2006). "Defining Ill-Defined Domains; A literature survey." In *Proc. of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains at the 8th International Conference on Intelligent Tutoring Systems*, V. Aleven, K. Ashley, C. Lynch, & N. Pinkwart (Eds.), NCU, Jhongli, Taiwan. 1-10.

[11] Mitchell Tom M. (1997). *Machine Learning*, McGraw Hill, Singapore.

[12] Pinkwart, N., Aleven, V., Ashley, K., & Lynch, C. (2007): "Evaluating Legal Argument Instruction with Graphical Representations using LARGO." To appear in *Proceedings of AIED 2007.*

[13] Robinson et al. (2006) "Increasing Text Comprehension and Graphical Note Taking using a Partial Graphical Organizer." *Journal of Educational research* 100(2) 103-111.

[14] Schworm, S.; Renkl, A. (2002) "Learning by solved example problems: Instructional explanations reduce self-explanation activity." In *Proc. of the Annual Conference of the Cognitive Science Society.* Lawrence Erlbaum, Mahwah, 816-821

[15] Van Gelder, T. (2002) "Argument Mapping with Reason!Able." *The American Philosophical Assoc. Newsletter on Philosophy and Computers.* 85-90.