Re-evaluating LARGO in the Classroom: Are Diagrams Better Than Text for Teaching Argumentation Skills?

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

[1]Clausthal University of Technology, Department of Informatics, Germany
[2]University of Pittsburgh, Intelligent Systems Program, Pittsburgh, PA, USA
[3]University of Pittsburgh, Learning Research and Development Center, Pittsburgh, PA, USA
[4]Carnegie Mellon University, HCI Institute, Pittsburgh, PA, USA

Abstract. Diagrams appear to be a convenient vehicle for teaching argumentation skills in ill-defined domains, but can an ITS provide useful feedback on students' argument diagrams without assuming a well-defined procedure for objectively evaluating argument? LARGO is an ITS for legal argumentation that supports students as they diagram transcripts of US Supreme Court oral argument. It provides on-demand advice by identifying small, interesting or incomplete patterns within students' graphs. We conducted a study in which LARGO was used as mandatory part of a first-year law school class. In contrast to prior findings in lab studies with voluntary participants, the use of LARGO did not lead to superior learning as compared to a text-based note-taking tool. These results can be partially attributed to low use of the graphical tools and advice by the students as well as (and possibly due to) a different motivational focus. Some evidence was found that higher engagement with the system led to better learning, leaving open the tantalizing possibility of helping especially lower-aptitude students through use of LARGO.

Keywords: Ill-defined Domains, Legal Argumentation, Diagram Representations, ITS Evaluation.

1 Introduction

In a variety of domains, a central goal of education is training students to produce *robust* arguments that not only address the current problem but survive the test of other examples and cases that have been encountered in the past or that may arise in the future. When a student proposes a rule for defining a class of mathematical objects, a theory for explaining scientific data, or a rule justifying a legal decision, one expects other students or the teacher to respond, "But what if…." That is, they test the proposal by posing hypothetical examples or cases that may occur and that highlight potential problems with the proposed rule or theory.

Law students are taught to make arguments through Socratic classroom dialogue, participation in moot court sessions and the analysis of examples, notably important precedents. These activities imitate court room arguments. Advocates before the court make their arguments by proposing *tests* or legal rules which, if adopted and used to

B. Woolf et al. (Eds.): ITS 2008, LNCS 5091, pp. 90–100, 2008.

decide the case at hand, would achieve their goals. To challenge these proposed decision rules, an opponent or judge may pose *hypothetical cases* that may occur, are relevant to the issues of the argument, and illustrate situations that the rule should cover but does not or decides wrongly given the underlying principles and policies of the law. The advocates can then respond by modifying their tests as needed to cover or avoid the hypothetical case, or by distinguishing the hypothetical situation from the facts of the case [12].

Interestingly, in legal education, teachers instruct students by engaging them in practice making and responding to such arguments, but seldom make explicit the process itself. If a student's argument has a flaw, the teacher does not explain the flaw; instead, the teacher typically will respond to the argument with a counterargument that exploits the flaw, thus leaving to the student the responsibility of later reflecting on why his argument was weak. It is not always clear why this approach is taken. It raises the possibility that students might learn better if their self-reflections about the process were explicitly guided. ITS systems such as CATO [1] and ArguMed [16] could help as tools both to give students practice in making arguments and to make explicit the process of argumentation.

Graphical representations of argument and argument diagramming have gained currency in recent years [4,13]. Proponents of argument diagrams argue that they can make the essential logical relations explicit while retaining formal validity. Work by Carr [5] in the legal domain indicated that the production of argument diagrams can improve students' ability to produce high-quality arguments, and Schank [14] showed that the production of diagrams can improve students' argument coherence. Recent work by Harrell [7] and Easterday et al. [6] has substantiated that argument diagrams can be useful learning tools. In summary, the current state of research suggests that diagrams are a useful educational tool, but controlled empirical studies are still rare.

The LARGO Intelligent Tutoring System [3,10,11] for legal argumentation supports students in the process of analyzing oral argument transcripts (taken from the U.S. Supreme Court). These are complex, real-world examples of the kind of Socratic arguing with tests and hypotheticals in which professors seek to engage students in class. However, they are written rather than purely oral as in the classroom, and thus may be good examples to use in reflecting upon the process of argument. Since these transcripts tend to be more complicated than classroom arguments, students probably need support in order to understand and reflect on them. LARGO provides that support by capitalizing on the pedagogical value of argument diagrams. While using the system, students read through the transcript and produce a graphical markup of it, identifying the key tests, hypotheticals, responses, and facts as well as the relationships between them. LARGO helps students by giving feedback in the form of self-explanation prompts.

In the fall of 2006, we conducted a study of LARGO with paid volunteers from the first year Legal Process course at the University of Pittsburgh's School of Law. The subjects analyzed a pair of cases using either LARGO or a text-based note-taking tool. We found no overriding differences between the two conditions. However, lower aptitude students, as measured by their Law School Admission Test (LSAT) score (a frequently-used predictor for success at law schools), showed higher learning gains using LARGO than using the note-taking tool. Also, the use of LARGO's on-demand help features was strongly correlated with learning [11]. Further analysis indicated that familiarity with the system led students to engage in better note taking [9].

Since participation in the study was voluntary, the students were self-selected (from among those enrolled in the course) for their interest in the curriculum, the ITS, and the pay. Many expressed an interest in the system, making it apparent that they were among the more inquisitive members of their class. We therefore concluded that a second study was necessary to further examine and substantiate the findings with non-voluntary participants. We sought out an opportunity where LARGO would be required in a course setting, so that we would have a sample of students that is more directly representative of the LARGO target population, and that (compared to our earlier study) may include a larger proportion of lower-LSAT students, for whom LARGO was most effective in that earlier study.

Based on our prior results, the two hypotheses for the new study are: a) Lower-aptitude students will derive more benefits from LARGO than their higher-aptitude peers, and b) additional experience with the system will improve students' use and benefit of it (i.e., we hypothesized stronger effects than in our previous study, if we include more study sessions). The following sections of this paper describe the type of argumentation LARGO teaches, and the design and results of the study.

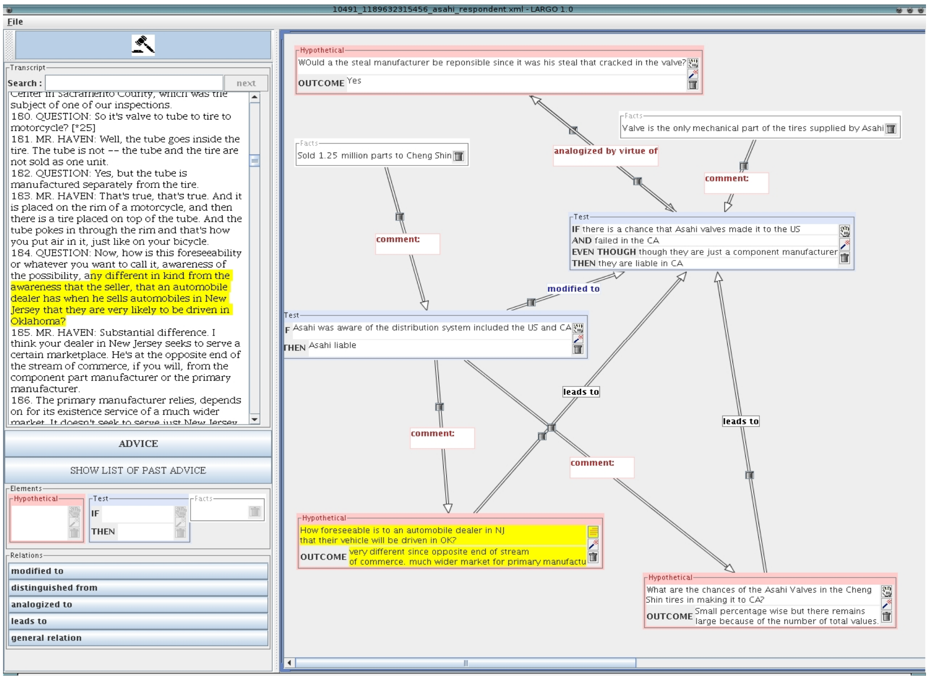
2 Arguing with Tests and Hypotheticals

An example taken from the case *Asahi Metal Industry Co. v. Superior Court,* (480 U.S. 102 (1987)), illustrates both the process taught, legal reasoning with test and hypotheticals, and the way in which LARGO's argument diagrams support learning. Law students encounter the *Asahi* case in their first semester "Legal Process" course. It deals with an important legal concept: *personal jurisdiction*, a court's power to require that a person appear in court and defend against a lawsuit.

Cases like *Asahi* involve a court in one state attempting to assert power over a non-resident of that state. In such cases, the principle that a state's courts may redress in-state harm conflicts with the U.S. Constitutional guarantee of "Due Process" requiring safeguards against the arbitrary exercise of government power. In *Asahi*, a motorcycle accident injured the driver and killed his wife. The driver filed a product liability claim against Cheng Shin, the Taiwanese maker of the tire in a California state court, alleging that a defect caused the accident. Shin in turn filed a claim against Asahi, the Japanese manufacturer of the tire's valve assembly, alleging that a defective valve caused the accident. Asahi moved to dismiss for lack of personal jurisdiction. The case made its way to the U.S. Supreme Court.

A typical Legal Process course book would include the Supreme Court's opinion in *Asahi* along with the facts of the case and its reasons. A law professor likely would engage the class in a Socratic discussion of the meaning and limitations of the Court's rule and alternate rules it might have adopted. If a student argued that Asahi should be subject to jurisdiction in California as its valves ended up there (i.e., proposed a test), the professor might ask: "How far up the stream of supply does it go? Does California have jurisdiction over the steel maker whose steel is in the valve?" (i.e., poses a hypothetical). Students learn to respond to such questions by analogizing the hypothetical to or distinguishing it from the case facts and defending the proposed rule, modifying the rule to accommodate it, or abandoning the rule in favor of another. In this way, the professor introduces students to the legal rules of personal jurisdiction, and to the *nature* of legal rules, the fact that they are defeasible, have an open texture, and may be applied differently in different circumstances.

At the U.S. Supreme Court, advocates often propose tests that decide the case at hand in their favor. The Justices often evaluate the tests by posing hypothetical cases like the one above to probe the test's meaning, its limits and consistency with precedents, principles, and policies. Thus, oral arguments at the U.S. Supreme Court provide complex examples of the kind of reasoning employed in the classroom, and therefore have a pedagogical value. Traditionally, however, oral arguments have not been employed in law school classes due to their complexity and lack of availability.



10491_1189632315456_asahi_respondent.xml - LARGO 1.0

Center in Sacramento County, which was the subject of one of our inspections.
180. QUESTION: So it's valve to tube to tire to motorcycle? [*25]
181. MR. HAVEN: Well, the tube goes inside the tire. The tube is not -- the tube and the tire are not sold as one unit.
182. QUESTION: Yes, but the tube is manufactured separately from the tire.
183. MR. HAVEN: That's true, that's true. And it is placed on the rim of a motorcycle, and then there is a tire placed on top of the tube. And the tube pokes in through the rim and that's how you put air in it, just like on your bicycle.
184. QUESTION: Now, how is this foreseeability or whatever you want to call it, awareness of the possibility, any different in kind from the awareness that the seller, that an automobile dealer has when he sells automobiles in New Jersey that they are very likely to be driven in Oklahoma?
185. MR. HAVEN: Substantial difference. I think your dealer in New Jersey seeks to serve a certain marketplace. He's at the opposite end of the stream of commerce, if you will, from the component part manufacturer or the primary manufacturer.
186. The primary manufacturer relies, depends on for its existence service of a much wider market. It doesn't seek to serve just New Jersey

Hypothetical
WOuld a the steal manufacturer be reponsible since it was his steal that cracked in the valve?
OUTCOME Yes

Facts
Valve is the only mechanical part of the tires supplied by Asahi

Facts
Sold 1.25 million parts to Cheng Shin

Test
IF there is a chance that Asahi valves made it to the US
AND failed in the CA
EVEN THOUGH though they are just a component manufacturer
THEN they are liable in CA

Test
F Asahi was aware of the distribution system included the US and CA
THEN Asahi liable

Hypothetical
How foreseeable is to an automobile dealer in NJ that their vehicle will be driven in OK?
OUTCOME very different since opposite end of stream of commerce. much wider market for primary manufactu

Hypothetical
What are the chances of the Asahi Valves in the Cheng Shin tires in making it to CA?
OUTCOME Small percentage wise but there remains large because of the number of total values.

Fig. 1. A student diagram for the Asahi argument

Figure 1 shows a student's actual LARGO diagram for a portion of the *Asahi* argument. The argument transcript is shown on the left, together with two buttons for advice and a palette from which the student can select the main graph elements. The student has identified two tests in the argument transcript, one a modification of the other, three hypotheticals posed, and a number of relations among them that (in this student's diagram somewhat imperfectly) reflect the role the hypotheticals play in evaluating the tests. LARGO helps students to find, diagram and relate the important elements of the text by providing hints based on small specific argument patterns [10].

3 Study Description

We carried out a study to evaluate LARGO within one section of the 2007 first-year Legal Process course at the University of Pittsburgh School of Law. All 85 subjects in the section were required to complete the activities in the study. The students were not paid but were given coffee gift cards as a token of appreciation. Since students are assigned randomly to one of three course sections, we have every reason to believe that the section that participated in the experiment is representative of their peers. The LARGO curriculum (which consisted of three personal jurisdiction cases) was integrated into the class as preparation for a graded writing assignment on personal jurisdiction, counting for 10% of their grade.

The students were assigned to two study conditions, balanced by LSAT scores, but otherwise assignment was random. The experimental group used a graphical version of LARGO that supported diagram creation and gave advice [11], as described above. The control group made use of a 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. During the second week they read background material on *Asahi* and annotated the transcript in LARGO or the text tool. They then answered two written questions about it without their diagrams or notes. Over the next two weeks they completed two more cases in the same way. 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 that had been used by the other condition, in order to compensate for any differences in learning between conditions prior to the course exam.

We classified the test items by type. Both the pre-test and post-test contained multiple-choice questions about: a) everyday reasoning with hypotheticals; b) generic aspects of tests and hypotheticals in legal argument; c) the domain of personal jurisdiction; and d) generic argument questions drawn from the LSAT. The post-test also contained: e) factual recall questions related to the specific transcripts studied during training; f) interpretation questions regarding these transcripts; and g) analysis and free-text questions regarding a novel case. We also grouped the items with respect to the aspect of the argument model to which they were most related (hypothetical, test, legal issues, legal policies, relation between test and hypothetical, response to hypothetical). The design of the study and the materials used were the same as in the 2006 study [11], except that one extra training session was added.

4 Results

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 conditions.

We excluded a total of 15 students from the analysis. Four candidly told us that they were not working and deliberately entered off-topic responses in the post-test. Two others completed the post-test in less than 30 minutes, less time than is needed merely to read the materials (approx. 50 minutes). The remaining 9 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 Control, 34 LARGO).

Table 1 contains the mean scores and standard deviations of the case-specific post-test questions (i.e., the post-test only items). Table 2 shows the pre-post gains for counterbalanced items shared between the tests. All scores are given on a [0,1] scale. Both tables show the results for all 70 students as well as the sub-results for the 27 low-LSAT students whose LSAT scores are below the median of 159. For this group, our previous study showed a positive effect of LARGO as compared to the text tool.

Table 1. Study 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 items	.63 (.09)	.64 (.09)	.64 (.08)	.61 (.11)
Case Interpretation	.46 (.11)	.48 (.10)	.45 (.10)	.49 (.11)
Case Recall	.71 (.10)	.73 (.12)	.73 (.09)	.67 (.14)
Hypotheticals	.71 (.12)	.71 (.14)	.72 (.13)	.64 (.14)
Legal issues	.39 (.49)	.35 (.48)	.50 (.52)	.38 (.51)
Legal policies	.36 (.49)	.29 (.46)	.50 (.52)	.23 (.44)
Relations tests/ hypotheticals	.48 (.11)	.50 (.11)	.49 (.11)	.48 (.16)
Responses to hypotheticals	.44 (.23)	.45 (.24)	.40 (.32)	.49 (.28)
Tests	.75 (.18)	.79 (.15)	.75 (.17)	.76 (.21)

Table 2. Study results for counterbalanced between tests

mean (sd) of gain score	All students (N=70)		Low-LSAT students (N=27)	
	Control	LARGO	Control	LARGO
All items	-0.01 (.16)	-0.04 (.18)	-0.01 (.13)	-0.08 (.19)
Everyday argumentation *	0.01 (.34)	-0.05 (.36)	0.09 (32)	-0.19 (.38)
Generic items	-0.01 (.31)	-0.01 (.27)	-0.02 (.28)	-0.03 (.25)
LSAT questions	-0.03 (.23)	-0.02 (.24)	-0.02 (.20)	-0.06 (.25)
Personal jurisdiction *	0.07 (.40)	-0.13 (.42)	0.00 (.35)	-0.21 (.32)
Hypotheticals	0.08 (.53)	0.00 (.49)	0.21 (.42)	0.00 (.41)
Relations tests/ hypotheticals	-0.01 (.22)	0.01 (.30)	-0.03 (.24)	-0.07 (.40)
Responses to hypotheticals	0.06 (.39)	0.01 (.34)	0.14 (.36)	-0.12 (.36)
Tests	-0.17 (.65)	-0.18 (.52)	-0.36 (.63)	-0.15 (.55)

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.

For the question types that were shared between pre-test and post-test (in a counterbalanced manner), the Control group gained significantly more than the LARGO group on the personal jurisdiction items (F(1,68) = 4.250; p<.05). For the low LSAT students, the Control group gained significantly more than the LARGO group on the "everyday hypothetical argumentation with hypotheticals" questions (F(1,25) = 4.313; p<.05). No other significant differences were found. A repeated measures analysis reveals that the only significant difference between pre-test and post-test scores is a drop for the low-LSAT LARGO students (F(1,10)= 5.333; p<.05) on the "personal jurisdiction" domain questions.

These results seemingly contradict our 2006 results where the low-LSAT LARGO students outperformed their Control peers on several important question types. When we analyzed the log files from the study sessions and the LARGO diagrams, we found that the 2006 students made far greater use of LARGO's advice functions than the students in the current study (see Table 3). Moreover, in the current study, the advice usage dropped over time unlike in 2006: during the last session, on average only 0.6 advice requests were made per case (1.6 during the first case). The diagrams created in the current study contained fewer elements and relations than those from the 2006 study, and students in the current study did not link their diagram elements to the transcript as often (31% vs. 87%).

Table 3. Advice usage and diagram complexity

mean (sd)	2006 study (N=15)	2007 study (N=34)
Clicks on Advice button (shows 3 hints) per case	10.1 (10.8)	1.8 (3.9)
Selection of one of the 3 shown hints per case	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
Number of elements in student graphs	9.6 (2.7)	7.5 (2.3)
Number of relations in student graphs	7.9 (2.3)	5.2 (2.9)
Rate of elements that are linked to the transcript	.87 (.23)	.31 (.31)

Table 4. Correlations between advice requests in LARGO and test scores

Pearson correlations	All students (N=34)			Low-LSAT students (N=13)		
	Pre-test	Post-test	Gain	Pre-test	Post-test	Gain
Case Interpretation	-	.03	-	-	.15	-
Case Recall	-	-.05	-	-	.02	-
Everyday argumentation	-.06	.34 *	.33	.07	.46	.29
Generic items	-.06	-.19	-.18	.06	-.14	-.21
LSAT questions	-.07	.02	.06	-.11	.30	.24
Personal jurisdiction	-.09	.21	.16	.04	.46	.30
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 *
Tests	.05	.06	-.03	-.07	.11	.16

*: significant correlations (p<.05).

Together, these results seem to indicate that LARGO's advice was a key factor in the positive effects that we observed in 2006, and that the graphical representation alone is not sufficient. We therefore analyzed if, within the current study, a higher number of advice requests correlates with higher post-test or gain scores.

The pre-test scores are not correlated with advice usage: students with higher pre-test scores did not use help more or less often than students with lower pre-test scores. However, advice usage is positively correlated with the post-test score for everyday argumentation items for all subjects. For low-LSAT students, the advice usage is also highly positively correlated to pre/post gains on items about responses to hypotheticals. The advice given by the system, apparently, helped these students to better understand how one can respond to a hypothetical during (legal or everyday) argument.

These strategies are mentioned in the feedback messages LARGO provides, which supports this hypothesis. Due to the relatively small number of low-LSAT LARGO students (N=13), additional correlations (e.g., for everyday argumentation or for personal jurisdiction) did not reach the level of statistical significance at the .05 level. However, the general trend is that advice seems to have a positive effect on the performance of the lower LSAT students.

5 Discussion

The study results did not confirm our initial hypotheses: the use of LARGO as a mandatory (though non-graded) part of a legal process course did not lead to learning gains when compared to a simple note-taking tool. Further, students in both conditions did not improve from pre-test to post-test in any of the tested categories even though they studied the materials for approximately 6 hours. These results are not in line with our 2006 findings with paid volunteers, even though the experiment was similar in all respects. We see three possible ways of accounting for these differences: student motivation, engagement with the system, and post-test design.

5.1 Motivational Issues

The extent to which users engage with a system depends on their specific goals. In 2006 the users were volunteers paid for their participation. As such they appear to have been more motivated to explore the system, to exercise key features such as graphical relations, links between diagram and transcript, and on-demand advice, and to take their time. Our present population comprised 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 tended to underutilize its key 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, depends on the extent to which users actually use these features. In our prior study, the low-LSAT students chose to make use of LARGO's key features and showed performance gains. In the present study, neither the high- nor the low-LSAT students did so consistently. Thus, the LARGO group derived fewer benefits from the system and performed no better than the Control group. To get students to engage with the beneficial features outside of the lab it seems necessary to better integrate the tool into the classroom. In the current study, use of LARGO was aligned with the course goals but not a core part of the course. Students were required to participate in the LARGO sessions, but were not graded on these activities. The payoff for the students lay in the preparation that the activities gave them for their future work. If we want the students to use the on-demand system functions, future studies of LARGO (and probably this result is valid also for other ITSs) should pay more nuanced attention to the specific motivation of the students, especially in real classroom situations. This can probably be done by assigning grades to the graphs that students create with LARGO and through in-class support (e.g., discussion of the benefits of LARGO for the learning goals).

5.2 Engagement with the System

Our analysis of the study data suggests that low use of the LARGO advice functions at least partially accounts for the lack of difference between the study conditions: the LARGO students who used the advice more frequently did better at some centrally important post-test questions. The low usage of important system features may be connected to motivational issues (cf. 5.1). Consequently, we may need to modify 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 in the diagram to specific passages in the transcript being studied, and 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 often do not ask for help even though they could benefit from it [2]. The diagrams the students created in the current study support this position. A large number of students' graphs had errors of a type that would be noted and commented on by LARGO if the student requested its advice. But since students did not do so very frequently, they were often not informed of their misconceptions.

How could LARGO be redesigned to avoid this problem? Presenting corrective feedback immediately after they make a mistake (as done by many successful ITS systems) would be problematic in the ill-defined domain of legal argumentation. As described in [10], LARGO's on-demand feedback avoids false error messages that are likely to occur in this domain, where it is often not clear whether a diagram correctly reflects an argument or not. False or inappropriate feedback would be very problematic also because the feedback LARGO gives is cognitively demanding (self-explanation prompts).

A reasonable alternative and a compromise between the two extremes, to be tested in further studies, could be to highlight diagram regions on which LARGO could give feedback (similar to the feedback in Andes [15]). Thus, students would be aware that feedback is available, but would not be forced to attend to it immediately (or at all). Another design option would be to structure the interaction with LARGO so that the students have "diagram creation" phases and also phases where they are explicitly asked to reflect on their diagrams, assisted by advice from LARGO.

Perhaps students could also be made to engage more with LARGO by requiring a clear and tangible "result" of their analysis (e.g., a "final test") that could be checked against what actually happened in court. Also, it may be interesting to give feedback to students indicating whether they did better or worse than the attorney in the actual case, or than peer students, and how their result relates to the final opinion of the court. Also, a future version of LARGO could present additional material not contained in the transcript, and engage students more in actually *making* arguments in addition to *analyzing* them.

5.3 Post-Test Design

Many researchers argue that even without feedback, diagrams are better than texts for learning argument skills. In that light, one would have expected a benefit of LARGO in this study even though the advice usage was low. However, this was not the case.

Could it be that the post-test somehow did not fully measure what was taught? At the content level, that notion can be rejected. The post-test items were well aligned with the tasks students had to solve in the training session. Yet, there was a subtle (and necessary) difference between what we tested and what was taught with LARGO. During training, the LARGO students created graphs, whereas the post-test employed a textual notation only, since this is the standard format in which legal argument and legal reasoning tasks are presented to students. However, the effectiveness of graphical tools generally strongly depends on the amount and type of usage of these tools [8], and our chosen format may have favored the students in the text condition The graphs created by the students could have been used for some of the questions on the posttests (and a few students asked for them for exactly that purpose), and would surely have helped, but we did not provide them. Thus, we tested whether training with graphs *transfers* to textual questions better than training with texts, not whether students were able to use the representations they created effectively in a post-test. As mentioned, we deemed a textual post-test to have higher ecological validity.

6 Summary and Conclusion

In this study we tested the LARGO ITS as a mandatory part of a first-semester law school course. Prior research on graphical argument representations has suggested that the graphical format of LARGO and the on-demand help it provides would be beneficial. However, 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 that graphs are better then text for learning complex argumentation skills was not confirmed. The students who used graphs were also no worse than the text users - since many ITSs for argumentation rely on the graph structure as a central component to enable the system feedback, this is still an important result for ITS designers. Yet, it contradicts our prior positive results with LARGO in lab studies [10].

Although we did not find a difference between the two conditions, the study provides some evidence that those students who engaged more with the graphs as evidenced by more frequent use of LARGO's advice function, especially the low-LSAT students, did better than the text condition. This finding is consistent with our 2006 study [10] in which the paid volunteers used more of the LARGO features and benefited from them.

One tentative conclusion to take away from this study is that graphs may still be better than text, but that engagement is essential. One way to support engagement could be to change the feedback mechanism. The current on-demand feedback is well suited for ill-defined domains since it avoids false error messages, but it remains to be explored whether prompting the student with messages (at the risk of giving inappropriate or suboptimal advice) or at least highlighting "weak regions" in diagrams will engage the students and not confuse them. Another take-home message of the study is that the subject's motivation is a decisive factor, especially when "leaving the lab" and entering the classroom with ITS technology. Apparently, and somewhat to our surprise, it can make a difference whether participation is voluntary or mandatory – and if it is mandatory, whether the students are motivated to participate in a manner so that the key ITS features are used, especially if their usage is on-demand. Future studies with LARGO – on its way toward regular classroom usage – will have to take these factors into account.

Acknowledgment. NSF Grant IIS-0412830 supported this work.

References

1. Aleven, V.: An intelligent learning environment for case-based argumentation. Technology, Instruction, Cognition, and Learning 4(2), 191–241 (2006)
2. Aleven, V., Stahl, E., Schworm, S., Fischer, F., Wallace, R.M.: Help Seeking in Interactive Learning Environments. Review of Educational Research 73(2), 277–320 (2003)
3. Ashley, K., Pinkwart, N., Lynch, C., Aleven, V.: Learning by Diagramming Supreme Court Oral Arguments. In: Proceedings of the 11th International Conference on Artificial Intelligence and Law, pp. 271–275. ACM Press, New York (2007)
4. van den Braak, S.W., van Oostendorp, H., Prakken, H., Vreeswijk, G.: A Critical Review of Argument Visualization Tools: do users become better reasoners? In: Workshop notes of the ECAI 2006; Workshop on Computational Models of Natural Argument, Italy (2006)
5. Carr, C.S.: Using Computer Supported Argument Visualization to Teach Legal Argumentation. In: Visualizing Argumentation, pp. 75–96. Springer, London (2003)
6. Easterday, M., Aleven, V., Scheines, R.: Tis better to construct than to receive? The effects of diagramming tools on causal reasoning. In: Proceedings of AIED, pp. 93–100. IOS Press, Amsterdam (2007)
7. Harrell, M.: The improvement of critical thinking skills in What Philosophy Is (Tech. Rep. CMU-PHIL-158). Carnegie Mellon University, Department of Philosophy (2004)
8. Hundhausen, C., Douglas, S., Stasko, J.: A Meta-Study of Algorithm Visualization Effectiveness. Journal of Visual Languages and Computing 13(3), 259–290 (2002)
9. Lynch, C., Ashley, K., Pinkwart, N., Aleven, V.: Argument diagramming as focusing device: does it scaffold reading? In: Proceedings of the AIED Workshop on Applications for Ill-Defined Domains, pp. 51–60 (2007)
10. Pinkwart, N., Aleven, V., Ashley, K., Lynch, C.: Toward Legal Argument Instruction with Graph Grammars and Collaborative Filtering Techniques. In: Ikeda, M., Ashley, K.D., Chan, T.-W. (eds.) ITS 2006. LNCS, vol. 4053, pp. 227–236. Springer, Heidelberg (2006)
11. Pinkwart, N., Aleven, V., Ashley, K., Lynch, C.: Evaluating Legal Argument Instruction with Graphical Representations Using LARGO. In: Proceedings of AIED, pp. 101–108. IOS Press, Amsterdam (2007)
12. Prettyman, E.B.: The Supreme Court's Use of Hypothetical Questions at Oral Argument. Catholic University Law Review 33, 555–591 (1984)
13. Reed, C., Walton, D., Macagno, F.: Argument Diagramming in Logic, Law and Artificial Intelligence. The Knowledge Engineering Review 22, 87–109 (2007)
14. Schank, P., Ranney, M.: Improved reasoning with Convince Me. Human Factors in Computing Systems. In: CHI 1995 Conference Companion, pp. 276–277. ACM, New York (1995)
15. VanLehn, K., Lynch, C., Schulze, K., Shapiro, J.A., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., Wintersgill, M.: The Andes Physics Tutoring System: Lessons Learned. International Journal of Artificial Intelligence and Education 15(3) (2005)
16. Verheij, B.: Artificial Argument Assistants for Defeasible Argumentation. Artificial Intelligence 150, 291–324 (2003)