Supporting Self-explanation of Argument Transcripts: Specific v. Generic Prompts

Vincent Aleven
Niels Pinkwart
Human-Computer Interaction Institute
Carnegie Mellon University
{aleven, nielsp}@cs.cmu.edu

Kevin Ashley
Collin Lynch
Learning Research and Development Center
Intelligent Systems Program
School of Law
University of Pittsburgh
ashley@pitt.edu, collinl@cs.pitt.edu

Abstract. We are developing an intelligent tutoring system that helps beginning law students learn argumentation skills through the study of transcripts of oral argument sessions before the US Supreme Court. These transcripts exemplify complex reasoning processes in which proposed decision rules are evaluated by holding them against real and hypothetical cases. As a first step, we investigated (without computer-based support) how to design good self-explanation prompts. In well-structured domains, *generic* prompts (e.g., "Explain.") may be most effective, because they leave students more latitude in discovering deficits in their own knowledge. However, in an ill-defined domain such as legal reasoning, *specific* prompts, which ask students to interpret a transcript in terms of a specific argumentation framework, may be more likely to help them arrive at insightful interpretations. In an experiment with 17 beginning law students, we found that the less able students (as measured by LSAT scores) learned better with specific prompts, as hypothesized, but the more able students learned better with generic prompts. This interaction was seen on test items that asked students to make arguments about a legal issue similar to that encountered in one of the transcripts. There was no significant interaction on items where students were asked to interpret a transcript dealing with a new area of the law (as opposed to making arguments). Thus, for less able learners in an ill-defined domain, the advantages of specific prompts outweigh those of generic prompts. It is surprising however how quickly the balance tips in favor of generic prompts. We are currently analyzing students' self-explanations to provide a deeper interpretation of the results.

Keywords: self-explanation, argumentation, legal reasoning, ill-defined domains

INTRODUCTION

We report on a project to develop an ITS for legal argumentation (e.g., Aleven, 2003; in press; Muntjewerff & Breuker, 2001). The legal domain is ill-structured in that cases present issues for which there seldom are uniquely right answers. Instead, reasonable arguments usually support competing answers, as evidenced by dissenting opinions eventually becoming the law of the land, by decisions reversed on appeal, and by a general reluctance of legal professionals to predict the outcome of legal cases. Competing reasons can be found in conflicting precedents, the sometimes ambiguous logical structure of statutes, and alternative interpretations of abstract, open-textured concepts in legal rules. This ill structure is unavoidable. (See, e.g., Frank, 1930; Llewellyn, 1951. But see Dworkin, 1986 for an argument that legal and moral questions do have right answers.) For instance, legislators write statutes in terms of abstract legal concepts to implement underlying legal policies, but they cannot foresee all of the scenarios to which a particular statute will be applied. Further, in real-world scenarios the policies often conflict, and subtle differences in facts can lead courts to resolve otherwise similar problems in different ways.

We focus on US Supreme Court oral argument, rapid-fire exchanges in which opposing attorneys propose decision rules to decide the case at hand (and cases like it), and the Justices explore the ramifications of these proposals by posing hypothetical fact situations and asking how they should be decided according to the proposed rules. Each side in the argument has one half hour to address the court; the Justices famously interrupt an advocate with questions. Arguing one's first case before the U.S. Supreme Court is a professional milestone –some experienced advocates become famous for their skills in making such arguments. Transcripts of these arguments have been published, and are readily available on-line through Westlaw[1]® and Lexis[2]®. Audio

[1] www.westlaw.com
[2] www.lexis.com

recordings of the proceedings are becoming increasingly available through websites like OYEZ[3]. We believe that these transcripts, due to their authenticity and high drama, will be motivating materials for beginning law students.

Our main goal is to help law students understand the kinds of argumentation processes that unfold in these transcripts and to develop some of the argumentation skills that are employed in these exchanges. Eventually, our goal is to develop an intelligent tutoring system that engages and guides students in this regard. Even if the transcripts are motivating, they are very challenging materials. They are different from most "worked-out examples" (e.g., Atkinson, Derry, Renkl, & Wortham, 2000) in that they show reasoning processes in their authentic raw form, complete with the false starts, blind alleys, and tangential lines of reasoning that are typically removed from annotated materials. The oral argument transcripts are messier; the Justices interrupt and advocates fumble to regroup.

As a stepping stone toward building an ITS, we study self-explanation, which has been shown to be an effective metacognitive strategy, although primarily in well-structured domains such as physics, biology, or geometry (Aleven & Koedinger, 2002; Chi, 2000; Renkl et al., 1997; but see Schworm & Renkl, 2002). While a number of cognitive science studies have produced evidence of the effectiveness of self-explanation prompts (Chi, de Leeuw, Chiu, & Lavancher, 1994; Renkl, 2002; Schworm & Renkl, 2002), we know of no studies that have asked specifically what kinds of prompts are the most effective. Many authors (e.g., Chi, 2000; VanLehn, Jones, & Chi, 1992) seem to have assumed that generic prompts (e.g., "explain this to yourself," where "this" refers to a line in a worked-out example or a sentence or paragraph of expository text) are the most effective, presumably because they increase the chances that individual students will be able to identify gaps in their own understanding, discover deficiencies in their mental models, or generate useful inferences. Specific prompts on the other hand, specific questions about how to interpret the materials, appear to have been considered less effective, at least in well-structured domains. It is possible that they would be more helpful in getting some students to realize that they have a gap in their understanding and may even hint at how to fill the gap (e.g., VanLehn et al., 1992). But specific prompts, which typically target a particular gap, are only likely to benefit those students who have that gap. For all other students, such prompts simply ask them to explain something that they understand already, which will not greatly impact their learning. It seems that it would be difficult to prompt all students for all gaps that they might possibly have. Even worse, specific prompts may rob students of the opportunity to make a range of useful inferences because such prompts, due to their specificity, focus their attention on one specific issue.

In an open-ended and ill-structured domain such as legal reasoning, however, the trade-off between specific and generic prompts may play out differently. A basic assumption of our work is that students will develop a better understanding of legal argumentation if they interpret it as a process of hypothesis formation and testing. We have designed an argumentation framework, based on that view, which is described below. Specific prompts that ask students to interpret the transcript in terms of this framework may be more beneficial than generic prompts that merely draw students' attention to particular passages, especially if students are unfamiliar with the framework. The prompts may spur useful inferences that students would not have made otherwise (e.g., Chi, 2000). In an open-ended domain, specific prompts may be helpful in a more general sense as well. In studying the transcripts, students may make many connections with prior knowledge and may generate many inferences regarding the issues that they read about. This assumption is reasonable in light of the fact that legal cases deal with real-life events. Further, many people have at least a basic understanding of what the law says in many areas and of the legal concepts being applied (e.g., "the right to privacy"). They are likely also to bring to bear their common sense notions of what is just. Thus, they may not experience discrete "gaps" or deficiencies in their knowledge or mental models, the way one would in a well-structured domain (e.g., VanLehn et al., 1992). For example, in mathematics or physics, if one does not see the relation between two equations, it may be harder to ignore the knowledge gap. In ill-structured domains, to the extent that there are such gaps, they may be "obscured" by the many inferences that can be made. Thus, in an ill-defined domain, specific prompts may provide just the right amount of focus: enough to give students a better idea of what inferences and interpretations are interesting, but not so much that they draw students' attention away from useful thoughts that they would otherwise have. In order to test this hypothesis, we conducted a small empirical study to compare the relative advantages of specific and general prompts for the study of US Supreme Court oral argument.

The paper is structured as follows: we first explain the framework for legal argumentation that we would like students to apply to the argumentation transcripts. We then describe the design and outcomes of the experiment, and discuss our results in light of literature on self explanation and ITSs for ill-structured domains.

A FRAMEWORK FOR INTERPRETING ARGUMENTATION TRANSCRIPTS

Our goal is to help students understand the normative and cognitive role of the Justices' hypotheticals in legal argument. As we noted above, advocates make their case by proposing a test or standard for deciding the issue at hand in this and future cases. These tests may be based on the relevant statutory or constitutional texts, if any, and interpretations in past cases involving the issue. The advocate asserts that (a) the proposed test or standard

[3] www.oyez.org/oyez/frontpage

is the right standard for the court to apply in deciding the issue, and (b) when applied to the facts of the case, the standard yields the outcome urged by the advocate. The Justices employ hypothetical cases to draw out the legal consequences of adopting the proposed standard and applying it to this and future cases. The hypotheticals explore the meaning of the proposed test, its consistency with relevant legal principles, policies, and past case decisions, its application to the case's facts, and its sensitivity to changes in the facts. In this work, we are trying to help students identify instantiations of a novel model of this kind of argumentation. In particular, we would like to help students identify (1) the proposed tests for deciding the current case and the reasons justifying it, (2) the hypotheticals that challenge the proposed test, the nature of the challenge, and the accompanying reasons, and (3) the advocate's response to the challenge in one of three forms: disputing the hypothetical's significance, modifying the proposed test, or abandoning the proposed test.

The targeted interpretative process is illustrated using the oral argument transcript in *Dennis LYNCH, etc., et al., Petitioners v. Daniel DONNELLY et al.* 465 U.S. 668 (1984), which was argued before the US Supreme Court on October 4, 1983. An excerpt appears in Table 1. The City of Pawtucket, Rhode Island erected an annual Christmas display in the heart of the shopping district. The display, which was owned by the city, comprised among other things, a Santa Claus house, reindeer pulling Santa's sleigh, candy-striped poles, a Christmas tree, carolers, cutout figures representing a clown, an elephant, and a teddy bear, hundreds of colored lights, and a large banner that reads "SEASONS GREETINGS." It also included a crèche consisting of the traditional figures, including the infant Jesus, Mary and Joseph, angels, shepherds, kings, and animals. Pawtucket residents and the American Civil Liberties Union filed suit, claiming that the city's inclusion of the crèche in the display was unconstitutional.

The Establishment Clause of the First Amendment to the U.S. Constitution states that "Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances." *U.S. Const. Amendment I.* The Supreme Court had explained that the purpose of the Establishment and Free Exercise Clauses of the First Amendment is "to prevent, as far as possible, the intrusion of either [the church or the state] into the precincts of the other." *Lemon v. Kurtzman,* 403 U.S. 602 (1971). At the same time, however, the Court had recognized that "total separation is not possible in an absolute sense. Some relationship between government and religious organizations is inevitable." *Ibid.* In every Establishment Clause case, the Court tries to reconcile the tension between the objective of preventing unnecessary intrusion of either the church or the state upon the other, and the reality that, as the Court has often noted, total separation of the two is not possible. Thus, the issue in the *Lynch v. Donnelly* case is whether Pawtucket, R.I.'s crèche display is a violation of the Establishment Clause. A small excerpt of the oral argument is shown in Table 1.

The attorney representing the ACLU argued that the Pawtucket Christmas display should be considered unconstitutional (line 139, Table 1, left column). As often happens in these transcripts, a test was implied although the advocate did not state one explicitly (e.g., there are no clearly marked "if" and "then" parts). One possible formulation of Mr. DeLuca's test is as follows: "if a city owns a fundamental religious symbol and displays it adjacent to its City Hall, it is in violation of the Establishment Clause." Other valid formulations may be possible, including formulations with more abstract or less abstract terms, and formulations with different numbers of pre-conditions, illustrating the ill-structured nature of the domain. One challenge that students face therefore is to recognize when advocates' statements imply a test and to arrive at a suitable and accurate formulation of that test. In response to the attorney's test, the Justices posed a hypothetical (in this case, a slight variation of the facts of the case – see line 141 in Table 1), in an apparent attempt to explore how low the attorney wanted to set the threshold for violations of the Establishment clause. The Justice's hypothetical was specifically aimed at exploring whether, under the attorney's proposed test, the display of a religious symbol adjacent to the City Hall is sufficient for the City to violate the Establishment Clause, even if the City does not own the symbol in question. The attorney's response (line 142) implies that ownership is not necessary and that mere sponsorship by the City of a display that contains a religious symbol, even one not owned by the City itself, is unconstitutional. This can be seen as a broadening of the test originally formulated. Once again the test is not stated in explicit "if-then" format, nor does the attorney indicate explicitly that he is changing his test, let alone how he is changing it. It is thus up to the student to provide an accurate formulation. As the example illustrates, relating a transcript to the argumentation model is an interpretative process that goes well beyond paraphrasing what occurs in the transcript. We are not claiming that this kind of detailed analysis of Supreme Court oral argument is necessary in order to fully understand the court's decision in the case. Rather, we mean to suggest that it is a viable and interesting way to learn about argumentation.

DESIGN OF THE EXPERIMENT

Materials

The materials we used in this study were transcripts of US Supreme Court oral arguments for two cases, including *Lynch v. Donnelley*, presented above. We edited the transcripts slightly, in order to reduce the time that students would need to work through them. However, as much as possible, we tried to retain their authentic nature. We then inserted self-explanation prompts into the transcripts, with "generic prompts" for the control group, and "specific prompts" for the experimental group. The prompts were inserted at places where we identified key components of our argumentation framework: tests, hypotheticals, and responses to hypotheticals. As we mentioned above, there is no one correct way of applying the "test/hypothetical/response" model to a given transcript – but for purposes of adding self-explanation prompts, agreement in this regard is not necessary. The specific prompts asked students to interpret the transcript in terms of our argumentation framework (see Table 1, column labeled "Specific SE prompt"). The generic prompts, inserted at the same locations in the transcripts, merely said "Explain." Not all contributions in the transcripts had associated prompts (Table 1 has a greater density of prompts than the overall transcript). The materials were presented to students as Excel spreadsheets.

Table 1. Excerpt of an argument transcript with examples of generic and specific self explanation prompts

Transcript	**Specific SE prompt**	**Generic SE prompt**
137. ORAL ARGUMENT OF AMATO A. DE LUCA, ESQUIRE ON BEHALF OF THE RESPONDENTS	Which party does Mr. DeLuca represent?	Explain.
138. MR. DE LUCA: Mr. Chief Justice, and may it please the Court, with the possible exception of the cross, the nativity scene is one of the most powerful religious symbols in this country, and most certainly one of the most powerful Christian religious symbols in this country. It is, as all of the parties agree and acknowledge, the biblical account of the birth of Christ, the Christian Messiah, the Redeemer, according to the gospels of Matthew and Luke as contained in the New Testament.		
139. Pawtucket's purchase, the maintenance, and the erection of the fundamental Christian symbol involves government in religion to a profound and substantial degree. It has aligned itself with a universally recognized religious symbol and belief. I would like to bring to the Court's attention that although the religious symbol, the creche, is contained in a display that is on private property -- it is owned by the Slater Museum Historical Society -- it is adjacent to the City Hall. City Hall is approximately 100 feet away from this area.	What is Mr. DeLuca's test concerning the issue of whether a city's creche display violates the Establishment Clause? Write as clear and succinct a version of his proposed test as you can.	Explain.
140. Also, the creche and the display itself is -- there is a ceremony that is held by the mayor of the city of Pawtucket each year, a lighting ceremony, which announces the commencement of the display in the Hodgson Park area. The music that is played at the display is the same music that is also played inside of City Hall, and all of the festivities that take place at the display and at City Hall are paid for and sponsored by the city of Pawtucket.	What is the significance of the proximity of the creche to City Hall?	Explain.
141. QUESTION: Well, Mr. DeLuca, you say that although the property, the real property, I take it, on which the creche is located is private, it is only -- it adjacent to city property. Now, if the city did not own the creche itself, so that everything that was contributed to the display, including the creche, were privately owned, it wouldn't violate the First Amendment, the fact that it was right next door to the City Hall, would it?	What is the relationship of the Justice's hypothetical to Mr. DeLuca's test?	Explain.
142. MR. DE LUCA: Well, I think that in the -- I think that in understanding that the city owns all of the symbols and all of the artifacts that are contained in this display, and assuming that that -- the creche were purchased and paid for privately without any other explanation that it is private, then I think it would still violate the establishment clause for the First Amendment, because there is no indication to anyone looking at that that the display or the creche is not part of the broader display which is put up and sponsored by the city.	How does Mr. DeLuca respond to the Justice's hypothetical? What effect would the response have on his proposed test? Explain whether the response to the hypothetical leads to a change in the proposed test, and if so, what change.	Explain.

Subjects

The 17 participants in the study were recruited from a group of students enrolled in a 6-week summer program for newly-accepted law students prior to their first year in law school. Students were selected for this program on the basis of such factors as extended time out of school, disadvantaged economic background, etc. Participation in the study was voluntary. All subjects were paid to participate in the experiment. The students were divided into two conditions, balanced in terms of LSAT scores. The LSAT is the Law School Admissions Test. It is a moderately good predictor of success in law schools, and many law schools in the US use LSAT scores as a factor in deciding which students to admit.

Procedure

All students participated in three sessions, each of which lasted approximately 2.5 hours. During the first two sessions, they studied transcripts of two US Supreme Court cases. At the beginning of each session, they were given a short introduction to the case they were about to study, some background material about the legal issues it presented, and a brief summary of the argumentation model. They then studied the transcript, typing answers to the self-explanation prompts into the Excel spreadsheet. Students in the "Specific" condition were given the transcripts with the specific prompts, students in the 'Generic" condition those with generic prompts. At the end of the first two sessions, all participants took a survey, but since the results are not yet available, we will not mention them any further. The third session was a post-test session, which consisted of two parts: an Argumentation Transfer Test and a Domain Transfer Test, described next.

Tests

In the Domain Transfer Test, students were given a transcript for a third case dealing with a different area of the law, compared to the first two cases. This time, the transcript did not contain any self-explanation prompts. Apart from that, the transcript was presented in the same format as before (i.e., Excel spreadsheet). The students then took a survey about this transcript, similar to the kind of survey they had completed at the end of each of the first two sessions, in which they were asked which of the Justices' hypotheticals was the most problematic and to assess the quality of the advocate's response and to formulate a better one if possible. This task was very similar in structure and materials (except for the absence of prompts and the switch to a new area of the law) to the tasks carried out during the first two sessions.

In the Argumentation Transfer Task, students were given a description of the facts of a case that dealt with a very similar legal issue to that encountered during the second session. This time, however, the students were not asked to study a transcript, but were asked to help an attorney prepare to argue the case before the US Supreme Court. Thus they were asked to formulate a test for deciding the case that would give a favorable outcome (as opposed to interpreting what test is being used in a transcript) and to predict what hypotheticals the Justices would be likely to pose (as opposed to interpreting what hypotheticals are being used in a transcript and why). Thus, they were engaged in *making* the kinds of arguments of which so far they had only studied examples.

Two legal writing instructors independently graded the surveys. Since there is no standard way of grading these kinds of surveys, we designed a grading form, which asked the grader to rate the quality of the tests, hypotheticals and responses formulated by the subjects, and to rate how well the subjects had understood the legal issues of the cases. Most items were graded using a 5-point Likert scale (e.g., "How well did the student formulate a test for Ms. Stone to propose that would lead to a favorable result for the ACLU?"). In addition, the form asked the graders to summarize or characterize the student's answers in textual form. The graders were also asked to rate the hypotheticals formulated by the students with respect to 12 (positive and negative) characteristics (e.g. "concise with no irrelevant details", "very creative", or "irrelevant to the argument").

RESULTS

We first evaluated the inter-rater reliability of the two legal writing instructors who graded the post-test materials. For the Likert Scale questions, which as mentioned cover the majority of the test items, we adjusted the grades assigned by one rater, subtracting one from each grade to achieve a common mean grade between the two graders. Then, counting as agreement grades that differ by no more than 1, we computed Cohen's Kappa as $\kappa=0.75$. This level indicates a satisfactory reliability in relative quality estimation. For the hypothetical characteristic questions, Cohen's Kappa was less than 0.7, but the percentage agreement was 0.74, which can be considered a satisfactory level of inter-rater reliability for yes/no questions. Having established that the inter-rater reliability was satisfactory, we based all of our following data analyses on the average of the two graders' opinions.

We first computed a single overall score for each subject, which included all survey items with Likert Scale or yes/no answers. We also computed subscores by only considering items that were related to a specific test (Argumentation Transfer / Domain Transfer) or a specific aspect of the argumentation model (test / hypothetical / response). With respect to the overall scores, in the full sample there was no main effect of condition, either on the Argumentation Transfer Test ($F(1,15)=0$, $p>.9$) or in the Domain Transfer Test ($F(1,15)=.725$, $p>.4$). Nor was there any significant difference with respect to the specific item types (or model aspects).

We then divided up the sample by means of a median split based on the students' LSAT scores, creating a. "lower LSAT" group that contained 8 students, and a "higher LSAT" group with 9 students. The students in the lower LSAT group scored significantly lower than their counterparts in the higher LSAT group throughout both tests ($F(1,15)=4.774$, $p<.05$), consistent with the predictive value claimed for the LSAT scores. We then considered whether the specific and generic prompts may have affected the students differently, depending on their ability level (as measured by LSAT scores). We found an interaction effect between ability level and condition, as illustrated in Figure 1: while the lower ability subjects group benefit more from specific prompts, the higher ability subjects are supported better by generic prompts. For the overall survey data, this interaction is at the borderline of significance ($p=.05$, repeated measures analysis). For the Argumentation Transfer Test considered separately, the interaction effect is statistically significant (F(3,13)=9.096, p<.01). (There was no statistically significant interaction on the Domain Transfer Test.) A more detailed analysis showed that this effect is largely due to test items in which students were asked to generate hypotheticals (interaction effect, F(3,13)=7.010, p<.01). For the test items that asked students to formulate a test, there is a marginally statistically interaction (F(3,13)=3.354, p<.1) indicating that the higher-ability subjects did better when trained with generic self-explanation prompts. In test items related to responses to hypotheticals, no significant interaction effect was found.

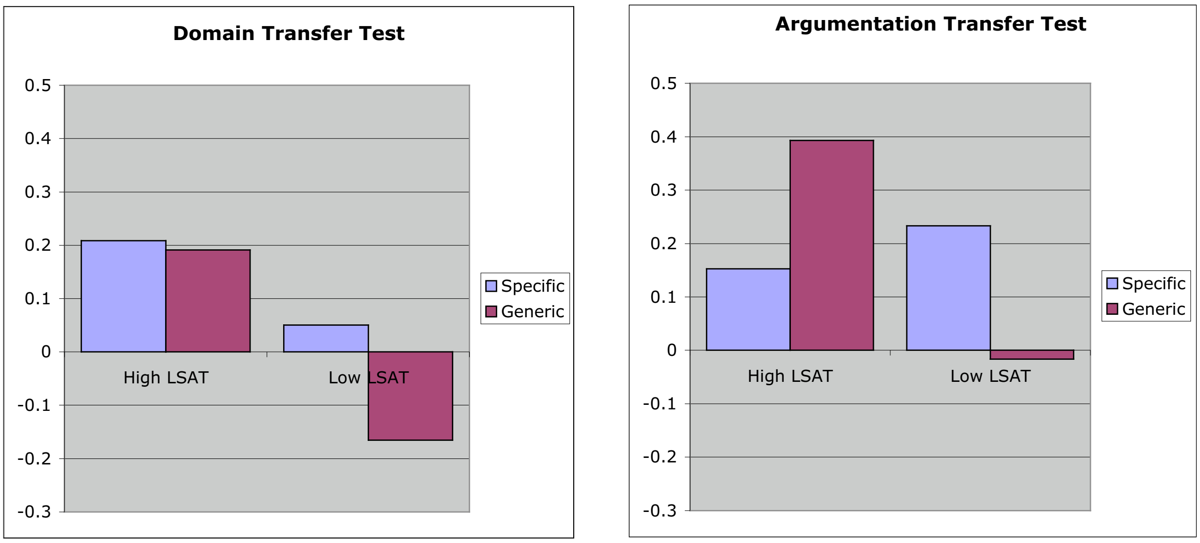

Figure 1. Results of the Domain / Argumentation Transfer Tests

We have begun to analyze the self-explanations in an attempt to better understand the interaction effect found at the post-test. The students' answers to one of the self-explanation prompts (i.e., the self-explanations typed during the first two sessions of the experiment – these self-explanations were part of the "training," not part of the post-test survey) illustrate that it is a challenging task to interpret a transcript and relate it to the argumentation framework described above. Table 2 shows the answers given by the 17 study subjects to the prompt shown in Table 1, line 139, where Mr. DeLuca, one of the attorneys, formulates a test, albeit implicitly so. The students' self-explanations are arranged according to their LSAT scores and the type of prompt that they received. The generic prompt, as always, merely asked students to "Explain." The specific prompt, shown in Table 1, asked students to provide a clear and succinct statement of Mr. DeLuca's test. As discussed above, one possible statement of the test is that "if a city owns a fundamental religious symbol and displays it adjacent to its City Hall, it is in violation of the Establishment Clause." No student provided a truly outstanding statement of the test. For example, it was rare to see an explicit if-then form. Some students did quite well (e.g., answers 3, 7, 9, and 14). Others provided statements that were not very "test-like," perhaps identifying some components of a test but not stating which way they cut (e.g., 1 and 5), or stating conditions that were rather abstract (e.g., 15). Some students did not really get to stating a test at all (e.g., 10 and 16, and obviously, 2). Thus, it is not an easy task to realize that a new test may be implied and to formulate the test, with a clear conclusion and a condition stated at an appropriate level of abstraction.

Based on the post-test results, one expects to see that the lower ability students provide better responses to the specific prompts than do they to the generic prompts, whereas for the higher ability students, one expects to see the opposite pattern. We will assess the explanations to see if this expectation is borne out. We will pay particular attention to whether students do "good things" in their responses (to either the specific or the generic prompts) that are not anticipated by the corresponding specific prompts. We will evaluate whether such "unprompted good things" are more likely to occur with generic prompts, and whether they relate to the relative advantages of specific and generic prompts mentioned in the introduction of the paper.

Table 2. Answers to specific and generic self explanation prompts.

	Answers to specific prompts	Answers to generic prompts
Lower LSAT	(1) Mr. DeLuca's test seems based upon whether the city pays for the display, what the meanings of the symbols are, where they are, and how they could be	(10) De Luca establishes that the creche has a religious legislative purpose that is excessive.
	(2) (none)	(11) Explains how the purchase of the creche by the city is definitely a government supporting a certain religious group by displaying the climax of their
	(3) DeLuca's test is that the RI city purchased, maintenance, and erected a Christian symbol. It clearly violates the Clause b/c governemnt is promotiong and subsidizing religion.	(12) Pawtucket's purchase, maintenance, and erection of the christian symbols involves the government in religion. The display is close to the the City Hall.
	(4) He sees the creche as a purely symbolic symbol and if the city displays and pays for it, it cannot say that it is not promoting religion. It cannot be seperated. Any government should not have religouis symbol on their property.	(13) De Luca is establishing his argument and trying to create a new context or framework for the hypotheticals to be drawn from. The basic argument is the religious symbol is universally recognized as a religious symbol and even though the creche is not on city property the creche is so close to city property it has the appearance of being part of the city's display.
Higher LSAT	(5) The amount of effort put out by the city in erecting the symbol. He also looks at where it is on display not just whether it is private or public property. It is really close to city hall.	(14) Council's test includes who purchased, maintained and displayed the creche. Furthermore, the test concludes that it is irrelevant that the creche was on private property. The display's close association with City Hall does not allow for patron's to distinguish what the city sponsors versus what is privately sponsored
	(6) It considers the degree of involvement of the government with religion. Is the government aligning itself with particular beliefs? How close is it to government property?	(15) Respondent begins argument by showing that the creche is known universally as a recognizable symbol, and that by displaying it, the city is promoting it. Respondent wants to point out that although the symbol has been placed on private land, it is close enough to the city hall to perhaps be confused as to being on city property
	(7) The city spent money on purchsing and maintaining the creche, a religious symbol, which satisfies the promotion of religion, which is a violation of the establishment clause. Also, though the creche is on private property it is in the backyeard of city hall. Through financial support the city has aligned itself	(16) He is trying to prove that the City has not fully seperated itself from the creche as it previously tried to convince the courts.
	(8) If the item is universally religious the government must not condone it. Here, Condoning means location of the nativity scene to the local government's Christmans celebrations.	(17) Making the point that the purchgase and matinance alone invlolves the govt. in the religion and helps to alighn them with that religion.
	(9) The involvement of the Government in religion requires the purchase, maintenance and facilitation of a fundamental religious symbol. The Government must have also aligned itself with that symbol and the belief. The symbol need not be on but near government	

DISCUSSION

Although the experiment did not confirm the hypothesis that specific prompts are more effective than generic prompts, it did produce an interesting result. There was no evidence in the full sample that students learn better when prompted with specific questions, rather than generic prompts that merely encourage them to explain. Instead, the experiment produced a statistically significant interaction, indicating that specific prompts are more helpful for students with lower ability, but generic prompts are more effective with better students. The interaction was seen on test items where students where asked to *make* arguments (as opposed to *studying* arguments, as they had done during the training phase) about a legal issue which by then had become somewhat familiar. There was no significant interaction on test items where students were asked to interpret a new transcript dealing with an unfamiliar legal issue. This interaction is consistent with the relative advantages and disadvantages of generic and specific prompts identified earlier in the paper. Specific prompts may be helpful because they have a scaffolding function: they lead students to useful inferences and perhaps lead them to identify gaps in their understanding (although the latter function is less certain in an ill-defined domain such as the current). However, with students who are inclined to make many inferences by themselves, without the help of a specific prompt, specific prompts may be harmful, in that they are likely to focus students' attention on a narrower set of inferences than they would otherwise have attended to. Generic prompts may be useful because they draw the students' attention to particular passages in the transcript, without restricting them to a small set of inferences.

At this point, it is not entirely clear to us what to make of the fact that an interaction effect was found with respect to the Argumentation Transfer Task but not with respect to the Domain Transfer Test. As mentioned, the latter involved a legal issue and area of the law that the students were not familiar with. It is possible that the new area was just too challenging for the students. It is possible also that having some basic grasp of the legal issues under study is a facilitating factor for learning argumentation skills. That interpretation is perhaps supported by the fact that during their summer program, outside of the study reported in this paper, the students had learned about the legal issues surrounding the First Amendment, which were targeted in the Argumentation Transfer Task but not the Domain Transfer Task. Further analysis of the data may shed more light on this issue.

In retrospect (although not *a priori),* what is surprising is not so much the fact that an interaction was found, but rather that it was found with a group of students very early on in their law school career – the study took place two months prior to the subjects' first year in law school. There was a significant range of student abilities in the sample, as measured by LSAT scores, although tilted somewhat towards the lower end of the LSAT spectrum. As mentioned, the participants in the study were recruited from the students enrolled in a summer school to help students prepare for law school. Participants in this program were selected based on factors such as extended time out of school and disadvantaged economic background. The fact that an interaction was found in this population suggests that the threshold ability level above which generic prompts are more effective is surprisingly low.

The findings from the current experiment are in line with findings by Conati and VanLehn (2000) who studied the effect of self-explanation support delivered by means of an intelligent tutoring system, and found that early on in students' development, more elaborate support is better, whereas later on, less elaborate support is better. The experiment dealt with worked-out physics problems (Newtonian mechanics), clearly a better-structured domain than argumentation. While the self-explanation support used in their experiment was more elaborate than in the current experiment, with the system dynamically selecting steps to explain based on a student model and providing feedback on students' self-explanations, their results could be interpreted (in tune with ours) as indicating that surprisingly early on in a student's development, support that is too elaborate becomes constraining. Coupling these results with literature on the expertise reversal effect, which states that in students' earlier developmental phases, examples are more effective than problem-solving practice, whereas in later phases the reverse is true (e.g., Kalyuga, Chandler, Tuovinen, & Sweller, 2001; Renkl & Atkinson, 2003) one gets an inkling then that the optimum level of support for any given student is continuously changing as the students develops. These changing needs present a challenge but also an opportunity for ITSs, suggesting that an ITS should be capable of varying its level of scaffolding even more so than previously thought (e.g., Collins, Brown, & Newman, 1989; VanLehn et al., 2000).

The current experiment seems to confirm some of the limitations of self-explanation prompts that were noted in previous experiments (Renkl et al., 1998) As illustrated, the responses that students typed to the self-explanation prompts leave room for improvement, consistent with Renkl's observations. Thus, another challenge for ITS research is to develop techniques for supporting self-explanation in an ill-defined domain beyond prompting, such as feedback on students' self-explanations (e.g., Aleven & Koedinger, 2002; Conati & VanLehn, 2000). The techniques developed in these earlier projects may not be applicable in ill-defined domain, since they depend on having an expert model that can produce a reasonably complete set of expert solutions. That assumption typically does not hold in an ill-defined domain, where often every (student or expert) solution is at least somewhat different (a point not mentioned in Herbert Simon's famous paper (1973) on ill-structuredness). In a companion paper to the current paper (Pinkwart, Aleven, Ashley, & Lynch, in press), we describe the next step in our project, the design of a system in which students self-explain argumentation transcripts by annotating them in a graphical language and receive feedback on their graphical annotations in the form of self-explanation prompts, as a form of adaptive prompting. The feedback is generated without the use of expert solutions.

CONCLUSION

In a well-structured domain, generic self-explanation prompts may be more effective than specific prompts, presumably because they leave individual students more latitude in discovering deficits in their own knowledge, even if specific prompts might provide more help in *leading* them toward specific deficits and possible ways of addressing them. We hypothesized that when students study complex, authentic argument transcripts in an ill-structured domain, specific prompts may provide useful scaffolding without being too constraining. We focused on prompts that ask students to interpret a transcript with respect to a specific argumentation framework and hypothesized that these prompts would lead students to useful inferences in a way that generic prompts would not. The results of the experiment indicate that this hypothesis holds true for lower ability students but not for higher ability students, who did better with generic prompts. The interaction was seen with respect to students' overall test scores, but was confined to an Argumentation Transfer Task, in which students were presented with a fact situation that involved a familiar legal issue, and were asked to make arguments rather than interpret arguments, as they had done during the training phase. To our knowledge, this interaction is a novel result in the self-explanation literature. The result seems consistent with the hypothesized advantages of specific prompts relative to generic prompts. The surprise is how low the threshold is in terms of the ability level at which the advantages of generic prompts outweigh those of specific prompts.

The aptitude-treatment interaction discovered in our experiment is relevant to the design of the next generation of adaptive ITS that engage students in self explanation. Self-explanation is an attractive educational approach for developing intelligent tutoring systems for ill-defined domains. Even without formal domain models, systems can prompt students to explain learning resources. However, if the interaction between ITS and learner is mediated through self-explanation prompts, the design of these prompts is essential. The current experiment suggests further that the prompts should be adapted to the student's ability level and that some amount of feedback on students' self-explanation is desirable.

An open issue is how the interaction effect that we found bears on current theories of self-explanation. We are currently coding the self-explanations given by the subjects in our study in order to analyze them for specific characteristics that might qualitatively explain the interaction.

ACKNOWLEDGEMENTS

This research is sponsored by NSF Award IIS-0412830. The contents of the paper are solely the responsibility of the authors and do not necessarily represent the official views of the NSF.

REFERENCES

Aleven, V. (in press). An intelligent learning environment for case-based argumentation. *Technology, Instruction, Cognition, and Learning.*

Aleven, V. (2003). Using background knowledge in case-based legal reasoning: a computational model and an intelligent learning environment. *Artificial Intelligence,* 150, 183-237.

Aleven, V., & Koedinger, K. R. (2002). An effective meta-cognitive strategy: learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science, 26*(2), 147-179.

Atkinson, R. K., Derry, S. J.; Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research 70*(2) 181-214.

Chi, M. T. H., de Leeuw, N., Chiu, M., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18,* 439-477.

Chi, M. 2000. Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In *Advances in Instructional Psychology*, 161-238. Hillsdale NJ, Lawrence Erlbaum.

Collins, A., Brown, J.S. & Newman, S.E. (1989). Cognitive apprenticeship: Teaching the craft of reading, writing and matematics. In L.B. Resnick (Ed.), Knowing, learning and instruction: Essays in honor of Robert Glaser (pp. 453-494). Hillsdale, NJ: Erlbaum.

Conati C. & VanLehn K. (2000). Toward computer-based support of meta-cognitive skills: a computational framework to coach self-explanation. *International Journal of Artificial Intelligence in Education, 11,* 398-415.

Dworkin, R. (1986). *Law's Empire.* Cambridge, MA: Harvard University Press.

Frank, J. (1930). *Law and the Modern Mind.* New York: Brentano's.

Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of Educational Psychology 93*(3), 579-588.

Llewellyn, K. (1951). *The Bramble Bush.* New York: Oceana.

Muntjewerff, A. J., & Breuker, J. A. (2001). Evaluating PROSA, a system to train solving legal cases. In J. D. Moore, C. L. Redfield, & W. L. (Eds.), *Johnson Artificial Intelligence in Education: AI-ED in the Wired and Wireless Future, Proceedings of AI-ED 2001* (pp. 278-285). Amsterdam: IOS Press.

Pinkwart, N., Aleven, V., Ashley, K., & Lynch, C. (in press). Toward legal argument instruction with graph grammars and collaborative filtering techniques. *Proceedings ITS 2006.*

Renkl, A., & Atkinson, R. K. (2003). Structuring the transition from example study to problem solving in cognitive skills acquisition: A cognitive load perspective. Educational Psychologist, 38, 15-22.

Renkl, A., Stark, R., Gruber, H., & Mandl, H. (1998). Learning from worked-out examples: the effects of example variability and elicited self-explanations. *Contemporary Educational Psychology, 23*, 90-108.

Renkl, A. (2002). Learning from worked-out examples: Instructional explanations supplement self-explanations. *Learning & Instruction, 12,* 529-556.

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

Simon, H. A. (1973). The structure of ill-structured problems. *Artificial Intelligence, 4,* 181-201.

VanLehn, K., Freedman, R., Jordan, P., Murray, C., Rosé, et al. (2000). Fading and deepening: The next steps for Andes and other model-tracing tutors. In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), *Intelligent Tutoring Systems: 5th International Conference* (pp. 474-483). Berlin: Springer-Verlag.

VanLehn, K., Jones, R. M., & Chi, M. T. H. (1992). A model of the self- explanation effect. *Journal of the Learning Sciences, 2*(1), 1-60.