Visualizing Arguments Improves Critical Thinking Skills

Maralee Harrell (mharrell@cmu.edu)
Carnegie Mellon University, Department of Philosophy, 135 Baker Hall
Pittsburgh, PA 15213 USA

I. Introduction

In the introductory philosophy class at Carnegie Mellon University, as at any school, one of the major learning goals is for the students the students to develop general critical thinking skills. There is, of course, a long history of interest in teaching students to “think critically” but it’s not always clear in what this ability consists. In addition, even though there are a few generally accepted measures (like the California Critical Thinking Skills Test, and the Watson Glaser Critical Thinking Appraisal, but see also Paul, et al., 1990 and Halpern, 1989), there is surprisingly little research on how sophisticated students’ critical thinking skills are or on the most effective methods for improving students’ critical thinking skills. The research that has been done shows that the population in general has very poor skills (Perkins, Allen & Hafner, 1983; Kuhn, 1991; Means & Voss, 1996), and that very few courses that advertise that they improve students’ skills actually do (Annis & Annis 1979; Pascarella, 1989; Stenning, Cox & Oberlander, 1995).

Most philosophers, and educators generally, can agree that one aspect of critical thinking is the ability to analyze, understand, and evaluate an argument. We believe that the students in our course do in fact gain this ability. This improvement, though, occurs in all the sections of our introductory class, which are taught by a variety of instructors using a range of instruction methods and so we are particularly interested in the efficacy of various alternative teaching methods to increase critical thinking performance.

What sorts of methods might be useful to aid students in analyzing and evaluating an argument? Larkin and Simon (1987) argue that diagrammatic representations of information can make recognition of important features and drawing inferences easier than a sentential representation of the same information. In addition, they argue that there are significant differences in the explicitness of the information as well as in the efficacy of search for information between diagrammatic and sentential representations of the same information. Winn (1991) makes a similar argument that maps and diagrams contain much more information that is easier to access than plain text just in virtue of the spatial relationships between the parts and between the parts and the frame. Indeed, research on student learning has consistently shown the efficacy of using diagrams to aid text comprehension (Armbruster & Anderson, 1984; Dansereau, et al.; Novak & Gowin, 1984; Schwartz & Rafael, 1985), as well as vocabulary development, postreading activities and writing preparation (Johnson, et al., 1986).

One candidate alternative teaching method, then, is instruction in the use of argument diagrams as an aid to argument comprehension (see Figure 1). We believe that the ability to construct argument diagrams significantly aids in understanding, analyzing, and evaluating arguments, both one’s own and those of others. If we think of an argument the way that philosophers and logicians do—as a series of statements in which one is the conclusion, and the others are premises supporting this conclusion—then an argument diagram is a visual representation of these statements and the inferential connections between them.

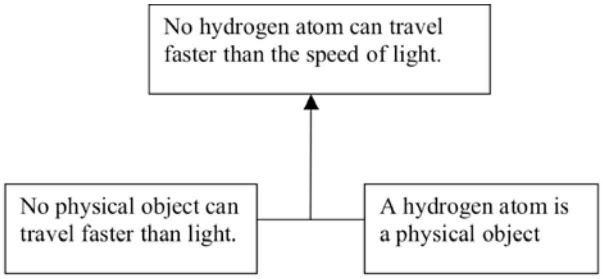
No hydrogen atom can travel faster than the speed of light.
No physical object can travel faster than light.
A hydrogen atom is a physical object

Figure 1: Example of a diagram for a simple argument.

Recent interest in argument visualization (particularly computer-supported argument visualization) has shown that the use of software programs specifically designed to help students construct argument diagrams can significantly improve students' critical thinking abilities over the course of a semester-long college-level course (Kirschner, Shum & Carr, 2003; Twardy, 2004; van Gelder, 2001, 2003). But, of course, one need not have computer software to construct an argument diagram; one needs only a pencil and paper. To our knowledge there has been no research done to determine whether the reported gains in critical thinking skills are due to the mere ability to construct argument diagrams, or the aid of a computer platform and tutor (or possibly both). We believe that it is the basic ability to construct argument diagrams—not the tools with which they are constructed—that is the important factor in the improvement of students' critical thinking skills.

Our hypothesis is that students who are able to construct argument diagrams and use them during argument analysis tasks will improve in performance on critical thinking tasks over the course of a semester long introductory philosophy class significantly more than students in the same class who do not have this ability.

Our introductory philosophy course was a natural place to study the skills acquisition of our students. We typically teach 4 or 5 lectures of this course each semester, with a different instructor for each lecture. While the general curriculum of the course is set, each instructor is given a great deal of freedom in executing this curriculum. For example, it is always a topics based course in which epistemology, metaphysics, and ethics are introduced with both historical and contemporary primary-source readings. It is up to the instructor however, to choose the order of the topics and the assignments. The students who take this course are a mix of all classes and all majors from each of the seven colleges across the University. This study tests this hypothesis by comparing the pre-test and post-test scores of students in the introductory philosophy class in the Spring and Fall of 2004 who were able to construct argument diagrams to the scores of those students in the introductory philosophy class who did not have this skill.

II. Method

A. Participants

One hundred thirty-nine students (46 women, 93 men) in each of the four lectures in the Spring of 2004, and 130 students (36 women, 94 men) in each of the five lectures in the Fall of 2004 of introductory philosophy (*80-100*) at Carnegie Mellon University were studied. Over the course of a semester, each lecture of the course had a different instructor and teaching assistant, and the students chose their section. In the Spring of 2004, there were 35 students (13 women, 22 men) in Lecture 1, 37 students (18 women, 19 men) in Lecture 2, 32 students (10 women, 22 men) in Lecture 3, and 35 students (5 women, 30 men) in Lecture 4. In the Fall of 2004, there were 24 students (6 women, 18 men) in Lecture 1, 36 students (6 women, 30 men) in Lecture 2, 26 students (9 women, 15 men) in Lecture 3, and 21 students (7 women, 14 men) in Lecture 4, and 23 students (8 women, 15 men) in Lecture 5. During the Spring 2004 semester, only the students in Lecture 1 (35 students) were taught the use of argument diagrams to analyze the arguments in the course reading, while the students in the other lectures (104 students) were taught more traditional methods of analyzing arguments. During the Fall 2004 Semester, the students in Lectures 1, 4 and 5 (68 students) were taught the use of argument diagrams to analyze the arguments in the course reading, while the students in Lectures 2 and 3 (62 students) were taught the more traditional method of analyzing arguments of listing the statements and identifying the conclusion.

During the Spring 2004 semester, only the students in Lecture 1 (35 students) were taught the use of argument diagrams to analyze the arguments in the course reading, while the students in the other lectures (104 students) were taught more traditional methods of analyzing arguments. During the Fall 2004 Semester, the students in Lectures 1, 4 and 5 (68 students) were taught the use of argument diagrams to analyze the arguments in the course reading, while the students in Lectures 2 and 3 (62 students) were taught the more traditional method of analyzing arguments of listing the statements and identifying the conclusion. The distribution of students who were taught and not taught the use of argument diagramming is given in Table 1.

Table 1: The distribution of students, and men and women who were taught and not taught argument diagramming in both Spring 2004 and Fall 2004

	Group	No. of Students	No. of Women	No. of Men
Spring 2004		139	46	93
	Taught Argument Diagramming	35	13	22
	Not Taught Argument Diagramming	104	33	71
Fall 2004		130	36	92
	Taught Argument Diagramming	68	21	47
	Not Taught Argument Diagramming	62	15	45

B. Materials

Prior to the first semester, the four instructors of the introductory philosophy course in the Spring of 2004 met to determine the learning goals of this course, and design an exam to test the students on relevant skills. In particular, the identified skills were to be able to, when reading an argument, (i) identify the conclusion and the premises; (ii) determine how the premises are supposed to support the conclusion; and (iii) evaluate the argument based on the truth of the premises and how well they support the conclusion.

We used this exam as the “pretest” (given in Appendix A)and created a companion “posttest” (given in Appendix B) for the Spring of 2004. For each question on the pre-test, there was a structurally (nearly) identical question with different content on the post-test. The tests each consisted of 6 questions, each of which asked the student to analyze a short argument. In questions 1 and 2, the student was only asked to state the conclusion (thesis) of the argument. Questions 3-6 each had five parts: (a) state the conclusion (thesis) of the argument; (b) state the premises (reasons) of the argument; (c) indicate (via multiple choice) how the premises are related; (d) the student was asked to provide a visual, graphical, schematic, or outlined representation of the argument; and (e) decide whether the argument is good or bad, and explain this decision. After a cursory analysis of the data from this first semester, we decided against including questions in which the student only had to state the conclusion. Thus, we designed a new pretest (given in Appendix C) and posttest (given in Appendix D) for the Fall of 2004, each of which consisted of five questions in which the student had again to analyze a short argument. Each question had the same five parts as the previous pretest and posttest.

C. Procedure

Each of the Lectures of 80-100 was a Monday/Wednesday/Friday class. In the Spring of 2004, the pretest was given to all students during the second day of class (i.e., Wednesday of the first week). The students in Lectures 1 and 4 were given the posttest as one part of their final exam (during exam week). The students in sections 2 and 3 were given the posttest on the last day of classes (i.e., the Friday before exam week). In the Fall of 2004, the pretest was given to all students during the third day of class (i.e., Friday of the first week), and the posttest on the last day of classes.

III. Results and Discussion

A. Test Coding

Pretests and posttests were paired by student, and single-test students were excluded from the sample. There were 139 pairs of tests for the Spring of 2004 and 130 pairs for the Fall of 2004. Tests which did not have pairs were used for coder-calibration, prior to each session of coding. The tests were coded during two separate sessions, using two different sets of coders: one session and set of coders for the Spring 2004 tests, and one for the Fall 2004. Each coder independently coded all pairs of tests in his or her group (278 total tests in Spring 2004, and 260 total tests in Fall 2004). Each pre-/post-test pair was assigned a unique ID, and the original tests were photocopied (twice, one for each coder) with the identifying information replaced by the ID. Prior to each coding session, we had an initial grader-calibration session in which the author and the two coders coded several of the unpaired tests, discussed our codes, and came to a consensus about each code. After this, each coder was given the two keys (one for the pre-test and one for the post-test) and the tests to be coded in a unique random order.

The codes assigned to each question (or part of a question, except for part (d)) were binary: a code of 1 for a correct answer, and a code of 0 for an incorrect answer. Part (e) of each question was assigned a code of “correct” if the student gave as reasons claims about support of premises for the conclusion and/or truth of the premises and conclusion. For part (d) of each question, answers were coded according to the type of representation used: Correct argument diagram, Incorrect or incomplete argument diagram, List, Translated into logical symbols like a proof, Venn diagram, Concept map, Schematic like: P1 + P2/Conclusion (C), Other or blank.

To determine inter-coder reliability, the Percentage Agreement (PA) as well as Cohen's Kappa (κ) and Krippendorff's Alpha (α) was calculated for each test (given in Table 2).

Table 2: Inter-coder Reliability: Percentage Agreement (PA), Cohen's Kappa (κ), and Krippendorff's Alpha (α) for each test

		PA	κ	α
Spring 2004	Pretest	0.85	0.68	0.68
	Posttest	0.85	0.55	0.54
Fall 2004	Pretest	0.88	0.75	0.75
	Posttest	0.89	0.76	0.76

As this table shows, the inter-coder reliability was fairly good. Upon closer examination, however, it was determined that, for each pair of coders, one had systematically higher standards than the other on the questions in which the assignment was open to some interpretation (questions 1 & 2, and parts (a), (b), and (e) of questions 3-6 for Spring 2004, and parts (a), (b), and (e) of questions 1-5 for Fall 2004). Specifically, for the Spring 2004 pretest, out of 385 question-parts on which the coders differed, 292 (75%) were cases in which Coder 1 coded the answer as "correct" while Coder 2 coded the answer as "incorrect"; and on the Spring 2004 posttest, out of 371 question-parts on which the coders differed, 333 (90%) were cases in which Coder 1 coded the answer as "correct" while Coder 2 coded the answer as "incorrect." Similarly, for the Fall 2004 pretest, out of the 323 question-parts on which the coders differed, 229 (77%) were cases in which Coder 1 coded the answer as "incorrect" while Coder 2 coded the answer as "correct"; and on the Fall 2004 posttest, out of 280 question-parts on which the coders differed, 191 (71%) were cases in which Coder 1 coded the answer as "incorrect" while Coder 2 coded the answer as "correct." In light of this, for each test, the codes from the two coders on these questions were averaged, allowing for a more nuanced scoring of each question than either coder alone could give.

Since we were interested in how the use of argument diagramming aided the student in answering each part of each question correctly, the code a student received for part (d) of each multi-part question (3-6 for Spring 2004 and 1-5 for Fall 2004) were preliminarily set aside, while the addition of the codes received on each of the other question-parts (questions 1 and 2, and parts (a), (b), (c), and (e) of questions 3-6 for Spring 2004 and parts (a), (b), (c), and (e) of questions 1-5 for Fall 2004) determined the raw score a student received on the test.

The primary variables of interest were the total pretest and posttest scores for the 18 question-parts for the Spring of 2004, and the 20 question-parts for Fall 2004 (expressed as a percentage correct of the equally weighted question-parts), and the individual average scores for each question on the pretest and the posttest. In addition, the following data was recorded for each student: which section the student was enrolled in, the student's final grade in the course, the student's year in school, the student's home college,[i] the student's sex, and whether the student had taken the concurrent honors course associated with the introductory course. Table 3 gives summary descriptions of these variables.

Table 3: The variables and their descriptions recorded for each student

Variable Name	Variable Description
Pre	Fractional score on the pre-test
Post	Fractional score on the post-test
Pre*	Averaged score (or code) on the pre-test for question *
Post*	Averaged score (or code) on the post-test for question *
Lecturer	Student's instructor
Sex	Student's sex
Honors	Enrollment in Honors course
Grade	Final grade in the course
Year	Year in school
College	Student's home college

B. Comparison of Gains by Diagram Use

Our main hypothesis is that the students who were able to construct correct argument diagrams would gain the most from pre-test to post-test. This implies that the number of correct argument diagrams a student constructed on the post-test was correlated to the student's gain from pretest to posttest. The straight gain, however, may not be fully informative if many students had fractional scores of close to 1 on the pretest. Thus, the hypothesis was also tested by determining the standardized gain: each student's gain as a fraction of what that student could have possibly gained.

Recall that for the Spring 2004 pretests and posttests, part (d) of questions 3-6 was coded based on the *type* of answer given. From this data, a new variable was defined that indicates how many correct argument diagrams a student had constructed on the posttest. This variable is PostCAD (value = 0, 1, 2, 3, 4). Similarly, for the Fall 2004 pretests and posttests, the type of answer given on part (d) of questions 1-5 was the data recorded. We again defined the variable PostCAD (value = 0, 1, 2, 3, 4, 5), indicating how many correct argument diagrams a student had constructed on the posttest.

For Spring 2004 there were very few students who constructed exactly 2 correct argument diagrams on the posttest, and still fewer who constructed exactly 4. Similar data obtained for Fall 2004. Thus, we grouped the students by whether they had constructed No Correct argument diagrams (PostCAD = 0), Few Correct argument diagrams (PostCAD = 1 or 2), or Many Correct argument diagrams (PostCAD = 3 or more). The results are given in Table 4 and in Figure 2.

Since the differences between No Correct and Few Correct is insignificant for both semesters, we did a planned comparison of the variables Post, Gain, and StGain for the group of Many Correct with the other two groups combined, again using the variable Pre as a covariate. This analysis indicates that the differences in the pretest scores was significant for predicting the gain (Spring 2004: $df = 1$, $F = 132.00$, $p < .001$; Fall 2004: $df = 1$, $F = 133.00$, $p < .001$), and the standardized gain (Spring 2004: $df = 1$, $F = 31.29$, $p < .001$; Fall 2004: $df = 1$, $F = 28.66$, $p < .001$). This analysis also indicates that in each semester, even accounting for differences in pretest score, the differences in the gains between students who constructed many correct argument diagram and the other groups were significant (Spring 2004: $df = 1$, $F = 28.13$, $p < .001$; Fall 2004: $df = 1$, $F = 37.78$, $p < .001$) as were the standardized gains (Spring 2004: $df = 1$, *F* = 22.27, *p* < .001; Fall 2004: *df* = 1, *F* = 34.14, *p* < .001), with the average gain and standardized gain being higher for those who constructed many correct argument diagrams than for those who did not.

Table 4: Mean gain and standardized gain for students who constructed No, Few, or Many Correct argument diagrams on the posttest.

		Gain		Standardized Gain	
		Mean	SD	Mean	SD
Spring 2004	No Correct	0.18	0.02	0.39	0.03
	Few Correct	0.17	0.02	0.37	0.04
	Many Correct	0.22	0.02	0.56	0.06
Fall 2004	No Correct	0.18	0.02	0.30	0.03
	Few Correct	0.19	0.03	0.27	0.04
	Many Correct	0.23	0.03	0.50	0.06

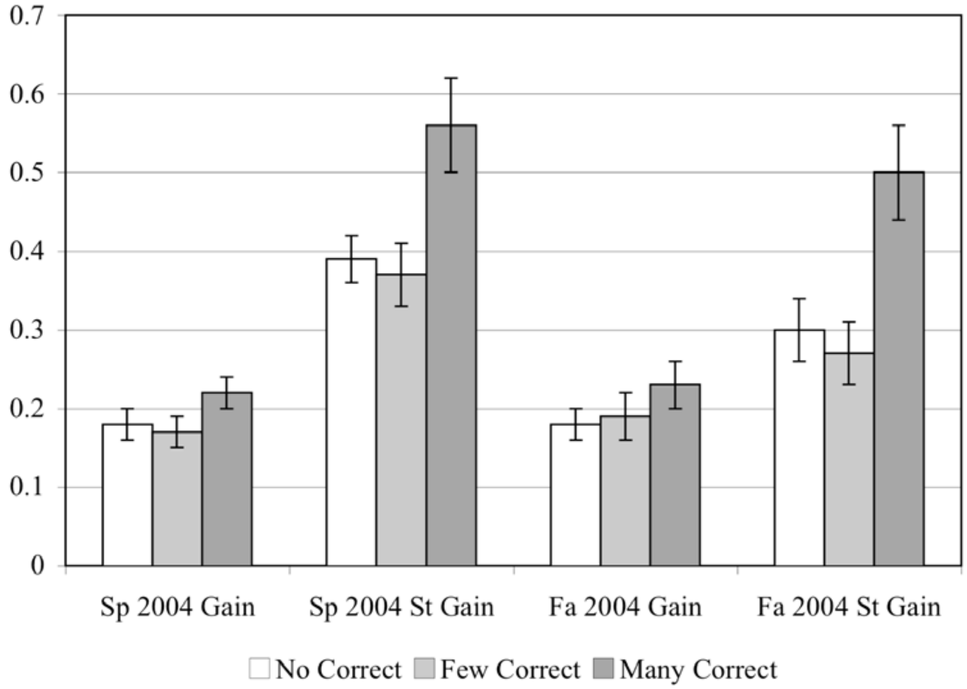

Figure 2: Comparison of gains and standardized gains in both Spring and Fall 2004 for students who constructed No, Few or Many Correct argument diagrams on the posttest.

These results show that the students who mastered the use of argument diagrams—those who constructed 3 or 4 correct argument diagrams on the posttest in the Spring of 2004, and those who constructed 3, 4 or 5 correct argument diagrams on the posttest in the Fall of 2004—gained the most from pretest to posttest, and gained the most as a fraction of the gain that was possible from pretest to posttest. Our hypothesis is thus highly confirmed.

Interestingly, those students who constructed few correct argument diagrams were roughly equal on all measures to those who constructed no correct argument diagrams. This may be explained by the fact that nearly all (85%) of the students who constructed few correct argument diagrams and all (100%) of the students who constructed no correct argument diagrams were enrolled in the lectures in which constructing argument diagrams was not explicitly taught; thus the majority of the students who constructed few correct argument diagrams may have done so by accident. This suggests some future work to determine how much the mere ability to construct argument diagrams aids in critical thinking skills compared to the ability to construct argument diagrams in addition to instruction on how to read, interpret, and use argument diagrams.

C. Prediction of Score on Individual Questions

The hypothesis that students who constructed correct argument diagrams improved their critical thinking skills the most was also tested on an even finer-grained scale by looking at the effect of (a) constructing the correct argument diagram on a particular question on the posttest on (b) the student's ability to answer the other parts of that question correctly. The hypothesis posits that the score a student received on each part of each question, as well as whether the student answered all the parts of each question correctly is positively correlated with whether the student constructed the correct argument diagram for that question.

To test this, a new set of variables were defined for each of the questions (3-6 for Spring 2004 and 1-5 for Fall 2004) that had value 1 if the student constructed the correct argument diagram on part (d) of the question, and 0 if the student constructed an incorrect argument diagram, or no argument diagram at all. In addition, another new set of variables was defined for each of the same questions that had value 1 if the student received codes of 1 for every part (a, b, c, and e), and 0 if the student did not. The histograms showing the comparison of the frequencies of answering each part of a question correctly given that the correct argument diagram was constructed to the frequencies of answering each part of a question correctly given that the correct argument diagram was not constructed are given in Figures 3 and 4.

We can see from the histograms that, on each question, those students who constructed the correct argument diagram were more likely—in some cases considerably more likely—to answer all the other parts of the question correctly than those who did not construct the correct argument diagram. Thus, these results further confirm our hypothesis: students who learned to construct argument diagrams were better able to answer questions that required particular critical thinking abilities than those who did not.

D. Prediction of Gain and Standardized Gain

While the results of the above sections seem to confirm our hypothesis that students who constructed correct argument diagrams improved their critical thinking skills more than those who did not, it is possible that there are many causes besides gaining diagramming skills that contributed to the students' improvement. In particular, since during both semesters the students of Lecturer 1 were the only ones explicitly taught the use of argument diagrams, and all of the students were able to chose their lecturer, it is possible that the use of argument diagrams was correlated with instructor's teaching ability, the student's year in school, etc.

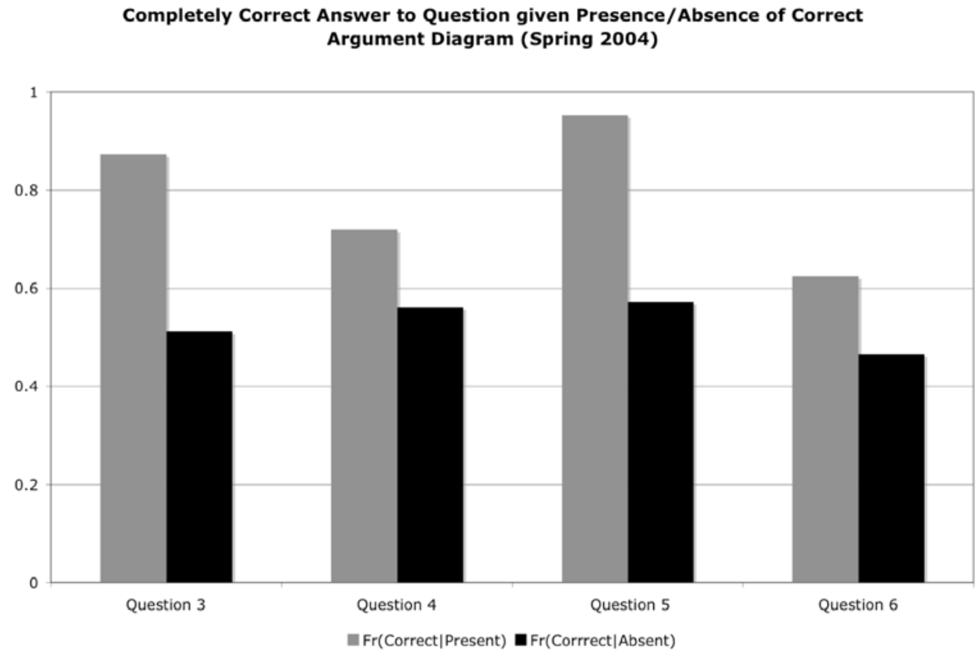

Figure 3: Histograms comparing the frequency of students (Spring 2004) who answered all parts of each question correctly given that they constructed the correct argument diagram for that question to the frequency of students who answered all parts of each question correctly given that they did not construct the correct argument diagram for that question.

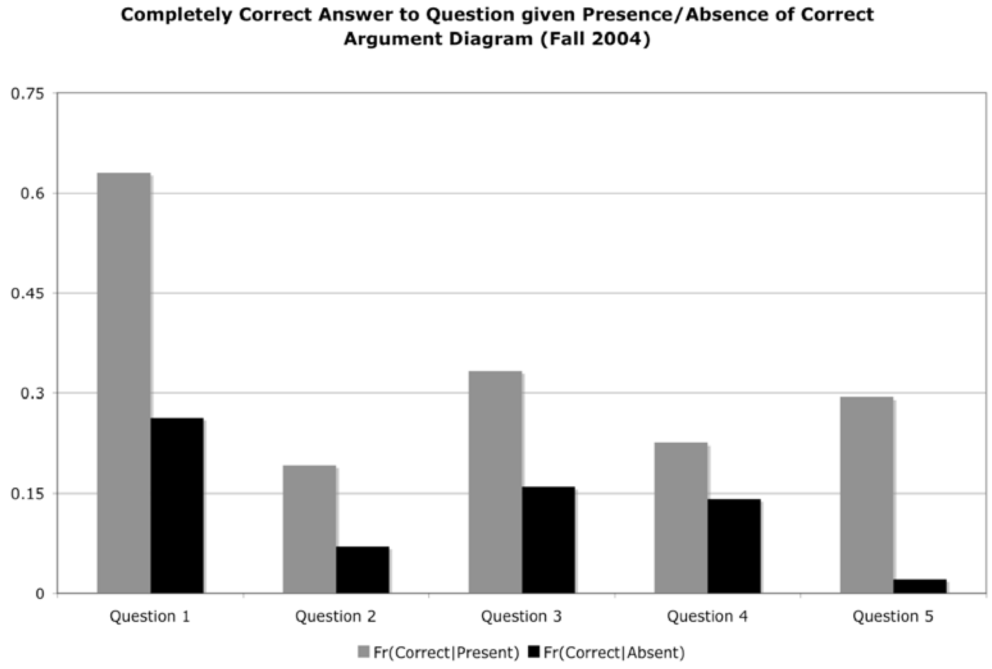

Figure 4: Histograms comparing the frequency of students (Fall 2004) who answered all parts of each question correctly given that they constructed the correct argument diagram for that question to the frequency of students who answered all parts of each question correctly given that they did not construct the correct argument diagram for that question.

To test the hypothesis that constructing correct argument diagrams was the only factor in improving students' critical thinking skills, we first considered how well we could predict the improvement based on the variables we had collected. We defined new variables for each lecturer that each had value 1 if the student was in the class with that lecturer, and 0 if the student was not (Lecturer 1, Lecturer 2, Lecturer 3, and Lecturer 4 for Spring 2004; and Lecturer 1, Lecturer 2, Lecturer 4, Lecturer 5, and Lecturer 6 for Fall 2004).

For each semester, we performed two linear regressions —one for the gain, and one for the standardized gain—using the pretest fractional score, the lecturer variables, and the variables Sex, Honors, Grade, Year and College as regressors. The results of these regressions showed that the variables Sex, Honors, Grade, Year and College are not significant as predictors in either semester of posttest score, gain or standardized gain. We then performed three more linear regressions on the data from each semester—again on the gain, and the standardized gain—this time using PostCAD as a regressor, in addition to the pretest fractional score, the lecturer variables, and the variables Sex, Honors, Grade, Year and College. Again, the results showed that the variables Sex, Honors, Grade, Year and College are not significant as predictors in either semester of posttest score, gain or standardized gain. Then, ignoring the variables that were not significant for either semester, we ran the regressions again. The results for the last set of regression analyses are given in Tables 5 and 6.

Table 5: Results of the regression analyses for each semester when the predicted variable is the gain.

	1st Regression		2nd Regression	
	Coeff.	SD	Coeff.	SD
Spring 2004				
Constant	0.534***	0.036	0.548***	0.035
Pretest	–0.694***	0.062	–0.756***	0.062
Lecturer 1	0.122***	0.025	0.052	0.031
Lecturer 2	0.071**	0.024	0.076**	0.023
Lecturer 3	0.080**	0.024	0.040	0.026
PostCAD			0.034**	0.010
Fall 2004				
Constant	0.505***	0.031	0.444***	0.030
Pretest	–0.657***	0.067	–0.788**	0.064
Lecturer 1	0.082*	0.039	0.074*	0.035
Lecturer 2	0.023	0.030	0.112***	0.031
Lecturer 5	–0.114***	0.032	–0.026	0.032
PostCAD			0.053***	0.009

Note: **p* < .05, ***p* < .01, ****p* < .001

Table 6: Results of the regression analyses for each semester when the predicted variable is the standardized gain.

	1st Regression		2nd Regression	
	Coeff.	SD	Coeff.	SD
Spring 2004				
Constant	0.818***	0.103	0.851***	0.101
Pretest	–0.948***	0.176	–1.096***	0.179
Lecturer 1	0.305***	0.069	0.136	0.090
Lecturer 2	0.199**	0.069	0.211**	0.067
Lecturer 3	0.209**	0.069	0.112	0.075
PostCAD			0.083**	0.029
Fall 2004				
Constant	0.623***	0.068	0.494***	0.065
Pretest	–0.659***	0.069	–0.951***	0.139
Lecturer 1	0.169*	0.084	0.150*	0.075
Lecturer 2	0.080	0.065	0.281***	0.067
Lecturer 5	–0.188**	0.069	–0.009	0.069
PostCAD			0.118***	0.020

Note: **p* < .05, ***p* < .01, ****p* < .001

The results show that in each set of regressions a student's pretest score was a highly significant predictor of the gain and standardized gain, which is to be expected, since the pretest score is an element of both. In addition, we can see that in each semester before including the variable PostCAD the coefficients for Lecturer 1 and Lecturer 3 were significantly positive for predicting a student's gain and standardized gain, while the coefficients for Lecturer 5 are significantly negative for predicting a student's gain and standardized gain. Interestingly, though, the coefficient for Lecturer 2 was significantly positive in the Spring of 2004, but insignificant in the Fall of 2004, for predicting a student's gain and standardized gain.

We can see further that in the Spring of 2004 when including the variable PostCAD, the variables Lecturer 1 and Lecturer 3 are no longer significant as predictors of gain and standardized gain; that is, when controlling for how many correct argument diagrams a student constructed, the students of Lecturers 1 and 3 were not significantly different from the students of Lecturer 4.

In the Fall of 2004, however, the coefficient of Lecturer 1 remains significantly positive as a predictor for gain and standardized gain when including the variable PostCAD; that is, even when controlling for how many correct argument diagrams a student constructed, the students of Lecturer 1 did better than the students of Lecturers 4, 5 and 6. Also in the Fall of 2004, after the variable PostCAD is introduced, the variable Lecturer 5 is no longer significant as a predictor of gain and standardized gain; that is, when controlling for how many correct argument diagrams a student constructed, the students of Lecturer 5 were not significantly different from the students of Lecturers 4.

Interestingly, the situation for Lecturer 2 is reversed; after introducing the variable PostCAD into the model in the Spring of 2004, the coefficient for Lecturer 2 was still significantly positive for predicting a student's gain and standardized gain, implying that when controlling for how many correct argument diagrams a student constructed, the students of Lecturer 2 did better than the students of the other lecturers. However, although Lecturer 2 had not been a significant predictor before the variable PostCAD was introduced in the Fall of 2004, after this variable is introduced the coefficient for Lecturer 2 becomes significantly positive for predicting gain and standardized gain, implying that when controlling for how many correct argument diagrams a student constructed, the students of Lecturer 2 did significantly better then the students of Lecturers 4, 5 and 6.

Importantly for testing our second hypothesis, in both semesters when PostCAD is introduced into the model, the coefficient for PostCAD is significantly positive for predicting a student's gain and standardized gain. For the Spring of 2004, this implies that the only measured factors that contributed to a student's posttest score and gain from pretest to posttest was being taught by Lecturer 2 and his or her ability to construct correct argument diagrams on the posttest. For the Fall of 2004, the analysis implies that the only measured factors that contributed to a student's posttest score and gain from pretest to posttest was being taught by Lecturer 1 or Lecturer 2 and his or her ability to construct correct argument diagrams on the posttest.

Thus, in the Spring of 2004, Lecturer 1—the only lecturer who explicitly taught argument diagramming in that semester—was not a direct contributing factor to the posttest score, gain or standardized gain. Rather, the students of Lecturer 1 did better only because they were significantly more likely than the other students to construct correct argument diagrams. However, in the Fall of 2004, Lecturer 1—only one of the lecturers who explicitly taught argument diagramming in that semester—is a direct contributing factor to the posttest score, gain and standardized gain. So, the students of Lecturer 1 performed as they did because they were both significantly more likely than the other students to construct correct argument diagrams, and benefited from other aspects of Lecturer 1's course.

IV. General Discussion

One set of skills we would like our students to acquire by the end of our introductory philosophy class can be loosely labeled "the ability to analyze an argument." This set of skills includes the ability to read a selection of prose, determine which statement is the conclusion and which statements are the premises, determine how the premises are supposed to support the conclusion, and evaluate the argument based on the truth of the premises and the quality of their support.

One purpose of argument diagrams is to aid students in each of these tasks. An argument diagram is a visualization of an argument that makes explicit which statement is the conclusion and which statements are the premises, as well as the inferential connections between the premises and the conclusion. Since an argument diagram contains only statements and inferential connections, it is clear which are the premises and which is the conclusion and how they are connected, and there is little ambiguity in deciding on what bases to evaluate the argument.

Since the scores on part (a) of each question were high on the pretest, and even higher on the posttest, it seems that the students taking *What Philosophy Is* at Carnegie Mellon University are already good at picking out the conclusion of an argument, even before taking this class. It also seems as though these students in general are *not* as able, before taking this class, to pick out the statements that served to support this conclusion, recognize how the statements were providing this support, and decide whether the support is good.

While on average all of the students in each of the sections improved their abilities on these tasks over the course of the semester, the most dramatic improvements were made by the students who demonstrated their ability to construct argument diagrams. Constructing the correct argument diagram was highly correlated in general with correctly picking out the premises, deciding how these premises are related to each other and the conclusion, and choosing the grounds on which to evaluate the argument.

It also seems that the access to a computer program that aids in the construction of an argument diagram (e.g. Reason!Able, Argutect, Inspiration) may not be nearly as important as the basic understanding of argument diagramming itself. The students who learned explicitly in class how to construct argument diagrams were all in section 1; these students saw examples of argument diagrams in class that were done by hand by the instructor, and they constructed argument diagrams by hand for homework assignments. While it may the case that access to specific computer software may enhance the ability to create argument diagrams, the results here clearly show that such access is not necessary for improving some basic critical thinking skills.

Interestingly, an analysis of the individual questions on the pretest yielded qualitatively similar results with respect to the value of being able to construct argument diagrams.

We conclude that taking Carnegie Mellon University's introductory philosophy course helps students develop certain critical thinking skills. We also conclude that learning how to construct argument diagrams significantly raises a student's ability to analyze, comprehend, and evaluate arguments.

V. Educational Importance of the Study

Many, if not most, undergraduate students never take a critical thinking course in their time in college. There may be several reasons for this: the classes are too hard to get into, the classes are not required, the classes do not exist, etc. It is difficult to understand, though, why any of these would be the case since the development of critical thinking skills are a part of the educational objectives of most universities and colleges, and since the possession of these skills is one of the most sought-after qualities in a job candidate in many fields.

What this study shows is that students can improve substantially their critical thinking skills if they are taught how to construct argument diagrams to aid in the understanding and evaluation of arguments. Although we studied only the effect of the use of argument diagrams in an introductory philosophy course, we see no reasons why this skill could not be used in courses in other disciplines. The creation of one's own arguments, as well as the analysis of others' arguments occurs in nearly every discipline, from Philosophy and Logic to English and History to Mathematics and Engineering (though, of course, the premises and acceptable evidence differ across domains). We believe that the use of argument diagrams would be helpful in any of these areas, both in developing general critical thinking skills, and developing discipline specific analytic abilities. We hope to perform more studies in the future to test these conjectures.

VI. Future Work

This study raises as many questions as it answers. While it is clear that the ability to construct argument diagrams significantly improves a student's critical thinking skills along the dimensions tested, it would be interesting to consider whether there are other skills that may usefully be labeled "critical thinking" that this ability may help to improve.

In addition, the arguments we used in testing our students were necessarily short and relatively simple. We would like to know what the effect of knowing how to construct an argument diagram would be on a student's ability to analyze longer and more complex arguments. We suspect that the longer and more complex the argument, the more argument diagramming would help.

It also seems to be the case that it is difficult for students to reason well about arguments in which they have a passionate belief in the truth or falsity of the conclusion (for religious, social, or any number of other reasons). We would like to know whether the ability to construct argument diagrams aids reasoning about these kinds of arguments, and whether the effect is more or less dramatic than the aid this ability offers to reasoning about less personal subjects.

In our classes at Carnegie Mellon University, we use argument diagramming not only to analyze the arguments of the philosophers we study, but also to aid the students with writing their own essays. We believe that, for the same reasons that constructing these diagrams helps students visually represent and thus understand better the structure of arguments they read, this would help the students understand, evaluate, and modify the structure of the arguments in their own essays better. We would like to know whether the ability to construct arguments actually does aid students' essay writing in these ways.

Lastly, unlike the relatively solitary activities in which students engage in our philosophy courses—like doing homework and writing essays—there are many venues in and out of the classroom in which students may engage in the analysis and evaluation of arguments in a group setting. These may include anything from classroom discussion of a particular author or topic, to group deliberations about for whom to vote or what public policy to implement. In any of these situations it seems as though it would be advantageous for all members of the group to be able to visually represent the structure of the arguments being considered. We would like to know whether knowing how to construct argument diagrams would aid groups in these situations.

REFERENCES

Annis, D., & Annis, L. (1979) Does philosophy improve critical thinking? *Teaching Philosophy*, 3, 145-152.

Armbruster, B.B., & Anderson, T.H. (1982). Mapping: Representing informative text graphically. In Holley, C.D. & Dansereau, D.F. (Eds.), *Spatial learning strategies*. New York: Academic Press.

Dansereau, D.F., Collins, K.W., McDonald, B.A., Holley, C.D., Garland, J., Diekhoff, G., & Evans, S.H. (1979). Development and evaluation of a learning strategy program. *Journal of Educational Psychology*, 71: 64-73.

Halpern, D.F. (1989). *Thought and knowledge: An introduction to critical thinking.* Hillsdale, NJ: L. Erlbaum Associates

Johnson, D.D., Pittleman, S.D., & Heimlich, J.E. (1986). Semantic mapping. *Reading Teacher*, 39: 778-783.

Kirschner, P.A., Shum, S.J.B., & Carr, C.S. (Eds.). (2003). *Visualizing argumentation: Software tools for collaborative and educational sense-making*. New York: Springer.

Kuhn, D. (1991). *The skills of argument*. Cambridge: Cambridge University Press.

Larkin, J.H., & Simon, H.A. (1987). Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cognitive Science*, 11, 65-99.

Means, M.L., & Voss, J.F. (1996). Who reasons well? Two studies of informal reasoning among children of different grade, ability, and knowledge levels. *Cognition and Instruction*, 14, 139-178.

Novak, J.D., & Gowin, D.B. (1984). *Learning how to learn*. New York: Cambridge University Press.

Pascarella, E. (1989). The development of critical thinking: Does college make a difference? *Journal of College Student Development*, 30, 19-26.

Paul, R., Binker., A., Jensen, K., & Kreklau, H. (1990). *Critical thinking handbook: A guide for remodeling lesson plans in language arts, social studies and science.* Rohnert Park, CA: Foundation for Critical Thinking.

Perkins, D.N., Allen, R., & Hafner, J. (1983). Difficulties in everyday reasoning. In W. Maxwell & J. Bruner (Eds.), *Thinking: The expanding frontier*. Philadelphia: The Franklin Institute Press.

Schwartz, R.M., & Raphael, T.E. (1985). Concept of definition: A key to improving students' vocabulary. *Reading Teacher*, 39: 198-205.

Stenning, K., Cox, R., & Oberlander, J. (1995). Contrasting the cognitive effects of graphical and sentential logic teaching: reasoning, representation and individual differences. *Language and Cognitive Processes*, 10, 333-354.

Twardy, C.R. (2004) Argument Maps Improve Critical Thinking. *Teaching Philosophy*, 27, 95-116.

van Gelder, T. (2001). How to improve critical thinking using educational technology. In G. Kennedy, M. Keppell, C. McNaught, & T. Petrovic (Eeds.), *Meeting at the crossroads: proceedings of the 18th annual conference of the Australian Society for computers in learning in tertiary education* (pp. 539-548). Melbourne: Biomedical Multimedia Uni, The University of Melbourne.

van Gelder, T. (2003). Enhancing deliberation through computer supported visualization. In P.A. Kirschner, S.J.B. Shum, & C.S. Carr (Eds.), *Visualizing argumentation: Software tools for collaborative and educational sense-making*. New York: Springer.

Winn, W. (1991). Learning from maps and diagrams. *Educational Psychology Review*, 3: 211-247.

Appendix A

80-100 Spring 2004 Pre-Test

A. Identify the conclusion (thesis) in the following arguments. Restate the conclusion in the space provided below.

1. Campaign reform is needed because many contributions to political campaigns are morally equivalent to bribes.

Conclusion:

2. In order for something to move, it must go from a place where it is to a place where it is not. However, since a thing is always where it is and is never where it is not, motion must not be possible.

Conclusion:

B. Consider the arguments on the following pages. For each argument:

(a) Identify the conclusion (thesis) of the argument.

(b) Identify the premises (reasons) given to support the conclusion. Restate the premises in the space provided below.

(c) Indicate how the premises are related. In particular, indicate whether they

(A) are each separate reasons to believe the conclusion,

(B) must be combined in order to provide support for the conclusion, or

(C) are related in a chain, with one premise being a reason to believe another.

(d) If you are able, provide a visual, graphical, schematic, or outlined representation of the argument.

(e) State whether it is a good argument, and explain why it is either good or bad. If it is a bad argument, state what needs to be changed to make it good.

3. America must reform its sagging educational system, assuming that Americans are unwilling to become a second rate force in the world economy. But I hope and trust that Americans are unwilling to accept second-rate status in the international economic scene. Accordingly, America must reform its sagging educational system.

(a) Conclusion:

(b) Premises:

(c) Relationship of the premises. Circle one: (A) (B) (C)

(d) Visual, graphical, schematic, or outlined representation of the argument:

(e) Good or bad argument? Why?

4. The dinosaurs could not have been cold-blooded reptiles. For, 2unlike modern reptiles and more like warm-blooded birds and mammals, some dinosaurs roamed the continental interiors in large migratory herds. In addition, the large carnivorous dinosaurs would have been too active and mobile had they been cold-blooded reptiles. As is indicated by the estimated predator-to-prey ratios, they also would have consumed too much for their body weight had they been cold-blooded animals.

(a) Conclusion:

(b) Premises:

(c) Relationship of the premises. Circle one: (A) (B) (C)

(d) Visual, graphical, schematic, or outlined representation of the argument:

(e) Good or bad argument? Why?

5. Either Boris drowned in the lake or he drowned in the ocean. But Boris has saltwater in his lungs, and if he has saltwater in his lungs, then he did not drown in the lake. So, Boris did not drown in the lake; he drowned in the ocean.
(a) Conclusion:
(b) Premises:
(c) Relationship of the premises. Circle one: (A) (B) (C)
(d) Visual, graphical, schematic, or outlined representation of the argument:
(e) Good or bad argument? Why?

6. Despite the fact that contraception is regarded as a blessing by most Americans, using contraceptives is immoral. For whatever is unnatural is immoral since God created and controls nature. And contraception is unnatural because it interferes with nature.
(a) Conclusion:
(b) Premises:
(c) Relationship of the premises. Circle one: (A) (B) (C)
(d) Visual, graphical, schematic, or outlined representation of the argument:
(e) Good or bad argument? Why?

Appendix B

80-100 Spring 2004 Final Exam

A. Identify the conclusion (thesis) in the following arguments. Restate the conclusion in the space provided below.

1. In spite of the fact that electrons are physical entities, they cannot be seen. For electrons are too small to deflect photons (light particles).
Conclusion:

2. Since major historical events cannot be repeated, historians are not scientists.]After all, the scientific method necessarily involves events (called "experiments") that can be repeated.
Conclusion:

B. Consider the arguments on the following pages. For each argument:
(a) Identify the conclusion (thesis) of the argument.
(b) Identify the premises (reasons) given to support the conclusion. Restate the premises in the space provided below.
(c) Indicate how the premises are related. In particular, indicate whether they
(A) are each separate reasons to believe the conclusion,
(B) must be combined in order to provide support for the conclusion, or
(C) are related in a chain, with one premise being a reason to believe another.
(d) Provide a visual, graphical, schematic, or outlined representation of the argument (for example, an argument diagram).
(e) State whether it is a good argument, and explain why it is either good or bad. If it is a bad argument, state what needs to be changed to make it good.

3. If species were natural kinds, then the binomials and other expressions that are used to refer to particular species could be eliminated in favor of predicates. However, the binomials and other expressions that are used to refer to particular species cannot be eliminated in favor of predicates. It follows that species are not natural kinds.
(a) Conclusion:
(b) Premises:
(c) Relationship of the premises. Circle one: (A) (B) (C)
(d) Visual, graphical, schematic, or outlined representation of the argument:
(e) Good or bad argument? Why?

4. Although Americans like to think they have interfered with other countries only to defend the downtrodden and helpless, there are undeniably aggressive episodes in American history. For example, the United States took Texas from Mexico by force. The United States seized Hawaii, Puerto Rico, and Guam. And in the first third of the 20th century, the United States intervened militarily in all of the following countries without being invited to do so: Cuba, Nicaragua, Guatemala, the Dominican Republic, Haiti, and Honduras.
(a) Conclusion:
(b) Premises:
(c) Relationship of the premises. Circle one: (A) (B) (C)
(d) Visual, graphical, schematic, or outlined representation of the argument:
(e) Good or bad argument? Why?

5. Either humans evolved from matter or humans have souls. Humans did evolve from matter, so humans do not have souls. But there is life after death only if humans have souls. Therefore, there is no life after death.
(a) Conclusion:
(b) Premises:
(c) Relationship of the premises. Circle one: (A) (B) (C)
(d) Visual, graphical, schematic, or outlined representation of the argument:
(e) Good or bad argument? Why?

6. Of course, of all the various kinds of artists, the fiction writer is most deviled by the public. Painters, and musicians are protected somewhat since they don't deal with what everyone knows about, but the fiction writer writes about life, and so anyone living considers himself an authority on it.
(a) Conclusion:
(b) Premises:
(c) Relationship of the premises. Circle one: (A) (B) (C)
(d) Visual, graphical, schematic, or outlined representation of the argument:
(e) Good or bad argument? Why?

Appendix C

80-100 Fall 2004 Pre-Test

Consider the following arguments. For each argument:
(a) Identify the conclusion (thesis) of the argument.
(b) Identify the premises (reasons) given to support the conclusion. Restate the premises in the space provided below.
(c) Indicate how the premises are related. In particular, indicate whether they
 (A) are each separate reasons to believe the conclusion,
 (B) must be combined in order to provide support for the conclusion, and/or
 (C) are related in a chain, with one premise being a reason to believe another.
(d) If you are able, provide a visual, graphical, schematic, or outlined representation of the argument.
(e) State whether it is a good argument, and explain why it is either good or bad. If it is a bad argument, state what needs to be changed to make it good.

1. Since major historical events cannot be repeated, historians are not scientists. After all, the scientific method necessarily involves events (called "experiments") that can be repeated.
(a) Conclusion:
(b) Premises:
(c) Relationship of the premises. Circle all that apply: (A) (B) (C)
(d) Visual, graphical, schematic, or outlined representation of the argument:
(e) Good or bad argument? Why?

2. The scientific method does not necessarily involve experimentation. For, if anything is a science, astronomy is. But the great cosmic events observed by astronomers cannot be repeated. And, of course, an experiment is, by definition, a repeatable event.
(a) Conclusion:
(b) Premises:
(c) Relationship of the premises. Circle all that apply: (A) (B) (C)
(d) Visual, graphical, schematic, or outlined representation of the argument:
(e) Good or bad argument? Why?

3. Although Americans like to think they have interfered with other countries only to defend the downtrodden and helpless, there are undeniably aggressive episodes in American history. For example, the United States to Texas from Mexico by force. The United States seized Hawaii, Puerto Rico, and Guam. And in the first third of the 20th century, the United States intervened militarily in all of the following countries without being invited to do so: Cuba, Nicaragua, Guatemala, the Dominican Republic, Haiti, and Honduras.
(a) Conclusion:
(b) Premises:
(c) Relationship of the premises. Circle all that apply: (A) (B) (C)
(d) Visual, graphical, schematic, or outlined representation of the argument:
(e) Good or bad argument? Why?

4. Politicians are forever attributing crime rates to policies—if the crime rates are decreasing, to their own policies; if the crime rates are increasing, to the "failed" policies of their opponents. But the fact is that crime rates are best explained in terms of demographics. For crime is primarily a young man's game. Whenever there is a relatively large number of young men between the ages of 15 and 30, the crime rates are high. And whenever this part of the population is relatively small, the crime rates are relatively low.
(a) Conclusion:
(b) Premises:
(c) Relationship of the premises. Circle all that apply: (A) (B) (C)
(d) Visual, graphical, schematic, or outlined representation of the argument:
(e) Good or bad argument? Why?

5. Small commercial fishing operations will continue to flourish only if restrictions on sport fishing are imposed. But the sport fishing lobby is powerful and vocal, for it is the sport of the rich and famous. And the sport fishing lobby does not want any restrictions. Consequently, restrictions on sport fishing activities are not likely in the near future. And, therefore, the small commercial fisherman is in big trouble.
(a) Conclusion:
(b) Premises:
(c) Relationship of the premises. Circle all that apply: (A) (B) (C)
(d) Visual, graphical, schematic, or outlined representation of the argument:
(e) Good or bad argument? Why?

Appendix D

80-100 Fall 2004 Final Exam

Consider the following arguments. For each argument:
(a) Identify the conclusion (thesis) of the argument.
(b) Identify the premises (reasons) given to support the conclusion. Restate the premises in the space provided below.
(c) Indicate how the premises are related. In particular, indicate whether they

(A) are each separate reasons to believe the conclusion,
(B) must be combined in order to provide support for the conclusion, and/or
(C) are related in a chain, with one premise being a reason to believe another.

(d) Provide a visual, graphical, schematic, or outlined representation of the argument (for example, an argument diagram).
(e) State whether it is a good argument, and explain why it is either good or bad. If it is a bad argument, state what needs to be changed to make it good.

1. No physical object can travel faster than light. A Hydrogen atom is a physical object, so no hydrogen atom can travel faster than the speed of light.
(a) Conclusion:
(b) Premises:
(c) Relationship of the premises. Circle all that apply: (A) (B) (C)
(d) Visual, graphical, schematic, or outlined representation of the argument:
(e) Good or bad argument? Why?

2. All brain events are physical events, and no physical events can be adequately accounted for in intensional terms, but it is only in terms of intensions that mental states can be adequately described. So, mental states cannot be brain events.
(a) Conclusion:
(b) Premises:
(c) Relationship of the premises. Circle all that apply: (A) (B) (C)
(d) Visual, graphical, schematic, or outlined representation of the argument:
(e) Good or bad argument? Why?

3. John and Robert Kennedy and Martin Luther King, Jr. were, like them or not, this country's last true national leaders. None of John Kennedy's successors in the White House has enjoyed the consensus he built, and everyone of them ran into trouble, of his own making, while in office. In the same way, none of this country's national spokespeople since Robert Kennedy and Dr. King has had the attention and respect they enjoyed.
(a) Conclusion:
(b) Premises:
(c) Relationship of the premises. Circle all that apply: (A) (B) (C)
(d) Visual, graphical, schematic, or outlined representation of the argument:
(e) Good or bad argument? Why?

4. The power set of any set (i.e. the set of all subsets of a given set) must be larger than the original set. The universal set is, by definition, the set of everything. Consequently, the universal set must not be possible, since its power set would have to contain more members than there are things in the universe.
(a) Conclusion:
(b) Premises:
(c) Relationship of the premises. Circle all that apply: (A) (B) (C)
(d) Visual, graphical, schematic, or outlined representation of the argument:
(e) Good or bad argument? Why?

5. Obviously, there is an objective moral law, for every sane person will agree that it is immoral to kill people at will. However, there is an objective moral law only if there is a moral Lawgiver who exists independently of human thinking. Hence, there is a moral Lawgiver who exists independently of human thinking. But God exists if there is a moral Lawgiver who exists independently of human thinking. Accordingly, God exists.
(a) Conclusion:
(b) Premises:
(c) Relationship of the premises. Circle all that apply: (A) (B) (C)
(d) Visual, graphical, schematic, or outlined representation of the argument:
(e) Good or bad argument? Why?

NOTES

[i] There are seven colleges at Carnegie Mellon in which undergraduate students may be enrolled: the College of Fine Arts (CFA), the Carnegie Institute of Technology (CIT), Carnegie Mellon University Honors College (CMU), the College of Humanities and Social Sciences (HSS), the Mellon College of Science (MCS), the School of Computer Science (SCS), the Tepper School of Business (TSB). The distribution of students in 80-100 from each college is given in Table A.

TABLE A
The distribution of home colleges in each lecture
in both Spring 2004 and Fall 2004

Lecture	**CFA**	**CIT**	**CMU**	**HSS**	**MCS**	**SCS**	**TSB**
Spring 2004 Total	5	40	7	48	12	15	12
Lecture 1	2	10	2	12	5	3	1
Lecture 2	2	5	3	8	4	6	7
Lecture 3	0	13	1	12	1	3	2
Lecture 4	1	12	1	16	2	3	2
Fall 2004 Total	3	37	6	44	18	9	13
Lecture 1	1	13	1	5	0	1	3
Lecture 2	0	6	1	20	3	5	1
Lecture 3	0	7	0	8	4	2	5
Lecture 4	2	5	2	4	7	0	1
Lecture 5	0	6	2	7	4	1	3