Using Argument Diagrams to Improve Critical Thinking Skills in 80-100 *What Philosophy Is*

Maralee Harrell

September 7, 2005

Technical Report No. CMU-PHIL-176

Philosophy

Methodology

Logic

Carnegie Mellon

Pittsburgh, Pennsylvania 15213

Using Argument Diagrams to Improve Critical Thinking Skills in 80-100 *What Philosophy Is*

Maralee Harrell[1]
Carnegie Mellon University

Abstract
After determining one set of skills that we hoped our students were learning in the introductory philosophy class at Carnegie Mellon University, we designed an experiment, performed twice over the course of two semesters, to test whether they were actually learning these skills. In addition, there were four different lectures of this course in the Spring of 2004, and five in the Fall of 2004; and the students of Lecturer 1 (in both semesters) were taught the material using argument diagrams as a tool to aid understanding and critical evaluation, while the other students were taught using more traditional methods. We were interested in whether this tool would help the students develop the skills we hoped they would master in this course. In each lecture, the students were given a pre-test at the beginning of the semester, and a structurally identical post-test at the end. We determined that the students did develop the skills in which we were interested over the course of the semester. We also determined that the students who were able to construct argument diagrams gained significantly more than the other students. We conclude that learning how to construct argument diagrams significantly improves a student's ability to analyze, comprehend, and evaluate arguments.

1. Introduction

In the introductory philosophy class at Carnegie Mellon University (*80-100 What Philosophy Is*), 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 (e.g. 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 the sophistication of students' critical thinking skills, or on the most effective methods for improving students' critical thinking skills. The research that has been done shows that the population of US college students in general has very poor skills (Perkins, et al., 1983; Kuhn, 1991; Means & Voss, 1996), and that very few college courses that advertise that they improve students' skills actually do (Annis & Annis 1979; Pascarella, 1989; Stenning et al., 1995).

Most philosophers can agree that one aspect of critical thinking is the ability to analyze, understand, and evaluate an argument. Our first hypothesis is that our students actually are improving their abilities on these tasks. We thus predict that students in the introductory philosophy course will exhibit significant improvement in critical thinking skills over the course of the semester. In addition to determining whether they are improving, though, we are

[1] I would like to thank Ryan Muldoon, Jim Soto, Mikel Negugogor, and Steve Kieffer for their work on coding the pre- and posttests; I would also like to thank Michele DiPietro, Marsha Lovett, Richard Scheines, and Teddy Seidenfeld for their help and advice with the data analysis; and I am deeply indebted to David Danks and Richard Scheines for detailed comments on many drafts.

particularly interested in the efficacy of various alternative teaching methods to increase critical thinking performance.

One candidate alternative teaching methods in which we are interested is instruction in the use of argument diagrams as an aid to argument comprehension. 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. For example, in the *Third Meditation*, Descartes argues that the idea of God is innate.

> It only remains to me to examine into the manner in which I have acquired this idea from God; for I have not received it through the senses, [since] it is never presented to me unexpectedly, as is usual with the ideas of sensible things when these things present themselves, or seem to present themselves, to the external organs of my senses; nor is it likewise a fiction of my mind, for it is not in my power to take from or add anything to it; and consequently the only alternative is that it is innate in me, just as the idea of myself is innate in me. (Descartes, 1641)

The argument presented here can be diagrammed as shown in Figure 1.

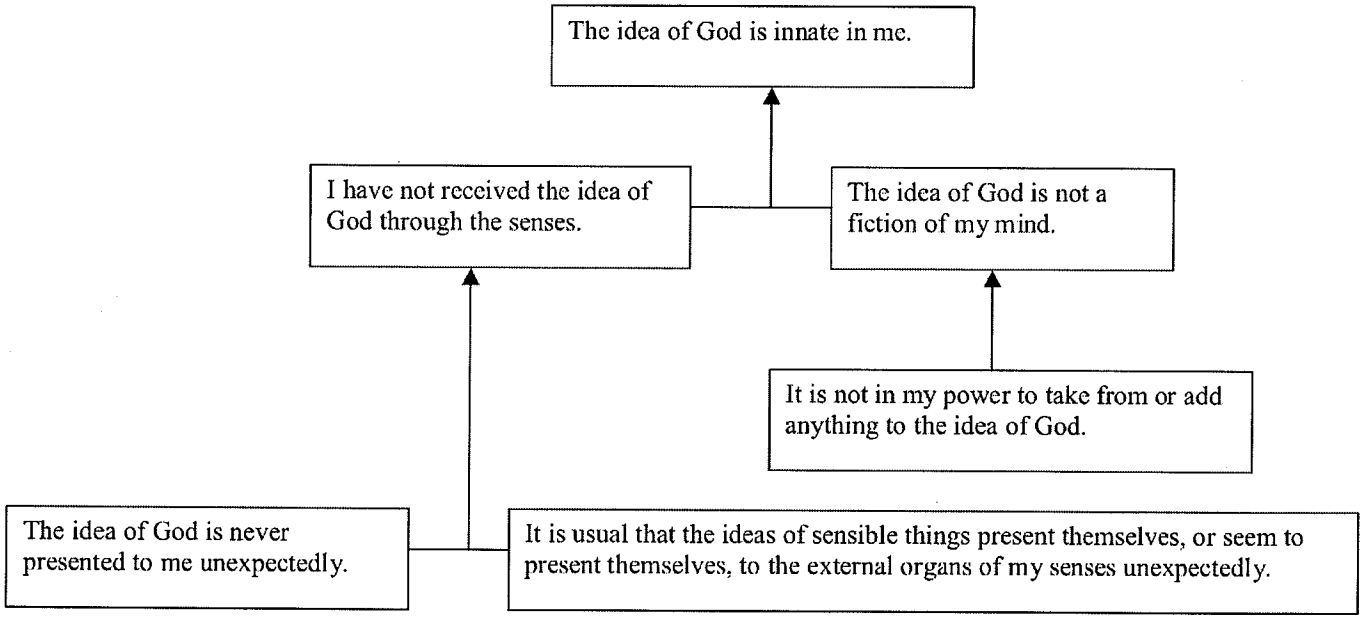

FIGURE 1 An argument diagram representing an argument in Descartes' *Third Meditation*.

Note not only that the text contains many more sentences than just the propositions that are part of the argument, but also that, proceeding necessarily linearly, the prose obscures the inferential structure of the argument. Thus anyone who wishes to understand and evaluate the argument may reasonably be confused. If, on the other hand, we are able to extract just the statements Descartes uses to support his conclusion, and visually represent the connections between these

statements, it is immediately clear how the argument is supposed to work and where we may critique or applaud it.

Recent research on 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, et al. 2003; 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 to determine whether the crucial factor is the mere ability to construct argument diagrams, or the aid of a computer platform and tutor, or possibly both.

Our second hypothesis is that it is the ability to construct argument diagrams that is the crucial factor in the improvement of students' critical thinking skills. This hypothesis implies that students who are taught how to construct argument diagrams and use them during argument analysis tasks should perform better on these tasks than students who do not have this ability. Carnegie Mellon University's introduction to philosophy course (*80-100 What Philosophy Is*), 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 a text, 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 second hypothesis by comparing the pretest and posttest scores of students in 80-100 in the Spring and Fall of 2004 who were taught how to use argument diagrams to the scores of those students in 80-100 who were not taught this skill.

2. Method

A. Participants

139 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 What Philosophy Is*) 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. Over both semesters there were 6 instructors, and 3 of those 6 (Lecturer 1, Lecturer 2 and Lecturer 4) taught a lecture in both semesters studied. During each semester, the students taught by Lecturer 1 were taught the use of argument diagrams to analyze the arguments in the course reading, while the students in the other lectures were taught more traditional methods of analyzing arguments. The distribution of instructors, students, men and women is given in Table 1.

TABLE 1
The distribution of instructors, students, men and women
in each lecture in both Spring 2004 and Fall 2004

Lecture	Instructor	No. of Students	No. of Women	No. of Men
Spring 2004	*(totals)*	*139*	*46*	*93*
Lecture 1	Lecturer 1	35	13	22
Lecture 2	Lecturer 2	37	18	19
Lecture 3	Lecturer 3	32	10	22
Lecture 4	Lecturer 4	35	5	30
Fall 2004	*(totals)*	*130*	*36*	*92*
Lecture 1	Lecturer 1	24	6	18
Lecture 2	Lecturer 2	36	6	30
Lecture 3	Lecturer 4	26	9	15
Lecture 4	Lecturer 5	21	7	14
Lecture 5	Lecturer 6	23	8	15

B. Materials

Prior to the first semester, the four instructors of 80-100 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 for the Fall of 2004 in which the student only had to state the conclusion (i.e. questions 1 and 2 from the Spring 2004 tests). Thus, we designed a new pretest (given in Appendix C) and posttest (given in Appendix D), each of which consisted of five questions in which the student had again to analyze a short argument. Each question in the Fall 2004 tests had the same five parts as questions 3-6 of the Spring 2004 tests. The Fall 2004 tests thus had 5 questions for directly testing critical thinking skills (rather than 4).

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.

3. 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
Pretest Spring 2004	0.85	0.68	0.68
Posttest Spring 2004	0.85	0.55	0.54
Pretest Fall 2004	0.88	0.75	0.75
Posttest Fall 2004	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,[1] 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. Average Gain from Pretest to Posttest for All Students

The first hypothesis was that the students' critical thinking skills improved over the course of the semester. This hypothesis was tested by determining whether the average gain of the students from pretest to posttest was significantly positive. 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. The mean scores on the pretest and the posttest, as well as the mean gain and standardized gain for the whole population of students for each semesters given in Table 4.

TABLE 4
Mean fractional score (standard deviation) for the pretest and the posttest, mean gain (standard deviation), and mean standardized gain (standard deviation)

	Pre	Post	Gain	StGain
Whole Population Spring 2004	0.59 (0.01)	0.78 (0.01)	0.19 (0.01)	0.43 (0.03)
Whole Population Fall 2004	0.46 (0.02)	0.66 (0.02)	0.20 (0.02)	0.34 (0.03)

For both Spring 2004 and Fall 2004, the difference in the means of the pretest and posttest scores was significant (paired t-test; $p < .001$), the mean gain was significantly different from zero (1-sample t-test; $p < .001$), and the mean standardized gain was significantly different from zero (1-sample t-test; $p < .001$). From these results we can see that our first hypothesis is confirmed: in each semester, overall the students did have significant gains and standardized gains from pretest to posttest.

C. Comparison of Gains of Students by Lecture and by Argument Diagram Use

Our second hypothesis was that the students who were able to construct correct argument diagrams would gain the most from pretest to posttest. Since the use of argument diagrams was only explicitly taught by Lecturer 1 each semester, we first tested this hypothesis by determining whether, in each semester, the average gain of the students taught by Lecturer 1 was significantly different from the average gain of the students in each of the other lectures. Again, though, the straight gain may not be fully informative if the mean on the pretest was not the same for each section, and if many students had fractional scores close to 1 on the pretest. Thus, we also tested this hypothesis using the standardized gain. The mean scores on the pretest and the posttest, as well as the mean gain and standardized gain, for the sub-populations of students in each lecture is given in Table 5 for the Spring 2004 data, and in Table 6 for the Fall 2004 data.

TABLE 5
Spring 2004: Mean fractional score (standard deviation) for the pretest and the posttest, mean gain (standard deviation), and mean standardized gain (standard deviation)

	Pre	Post	Gain	StGain
Lecturer 1	0.64 (0.02)	0.85 (0.02)	0.21 (0.02)	0.51 (0.07)
Lecturer 2	0.63 (0.02)	0.80 (0.02)	0.17 (0.02)	0.42 (0.05)
Lecturer 3	0.58 (0.02)	0.79 (0.01)	0.21 (0.02)	0.48 (0.04)
Lecturer 4	0.53 (0.03)	0.70 (0.02)	0.17 (0.03)	0.32 (0.05)

TABLE 6
Fall 2004: Mean fractional score (standard deviation) for the pretest and the posttest, mean gain (standard deviation), and mean standardized gain (standard deviation)

	Pre	Post	Gain	StGain
Lecturer 1	0.68 (0.04)	0.82 (0.02)	0.14 (0.03)	0.35 (0.09)
Lecturer 2	0.50 (0.02)	0.70 (0.02)	0.20 (0.03)	0.38 (0.05)
Lecturer 4	0.28 (0.03)	0.62 (0.02)	0.34 (0.04)	0.45 (0.04)
Lecturer 5	0.35 (0.03)	0.51 (0.03)	0.16 (0.03)	0.21 (0.06)
Lecturer 6	0.47 (0.04)	0.64 (0.04)	0.18 (0.04)	0.32 (0.06)

Since there was such variability in the scores on the pretest among the different lecturers in each semester, we ran an ANCOVA on the each of the variables Post, Gain, and StGain, with the variable Pre used as the covariate. This analysis indicates that in both semesters, the differences in the pretest scores was significant for predicting the posttest scores (Spring 2004: $df = 1, F = 24.36, p < .001$; Fall 2004: $df = 1, F = 27.25, p < .001$), the gain (Spring 2004: $df = 1, F = 125.50, p < .001$; Fall 2004: $df = 1, F = 79.30, p < .001$), and the standardized gain (Spring 2004: $df = 1, F = 29.14, p < .001$; Fall 2004: $df = 1, F = 18.06, p < .001$).

In addition, this analysis indicates that for both semesters, even accounting for differences in pretest score, the differences in the posttest scores among the lecturers were significant (Spring 2004: $df = 3, F = 8.71, p < .001$; Fall 2004: $df = 4, F = 6.53, p < .001$), as were the differences in the gains (Spring 2004: $df = 3, F = 8.71, p < .001$; Fall 2004: $df = 4, F = 6.53, p < .001$) and the standardized gains (Spring 2004: $df = 3, F = 6.84, p < .001$; Fall 2004: $df = 4, F = 4.34, p < .001$).

This analysis shows that a student's lecturer is a significant predictor of posttest score, gain, and standardized gain, but it does not tell us how the lecturers are different. The hypothesis is that the posttest score, gain and standardized gain for students of Lecturer 1 is significantly higher than for all the other lecturers. Thus, we did a planned comparison of the variables Post, Gain, and StGain for Lecturer 1 with the other lecturers combined, again using the variable Pre as a covariate. This analysis again indicates that, for both semesters, the differences in the pretest scores was significant for predicting the posttest scores (Spring 2004: $df = 1, F = 32.28, p < .001$; Fall 2004: $df = 1, F = 36.96, p < .001$), the gain (Spring 2004: $df = 1, F = 107.37, p < .001$; Fall 2004: $df = 1, F = 79.24, p < .001$), and the standardized gain (Spring 2004: $df = 1, F = 21.42, p < .001$; Fall 2004: $df = 1, F = 13.20, p < .001$).

In addition, this analysis indicates that for both semesters, even controlling for differences in pretest score, the differences in the posttest scores between the students of Lecturer 1 and the other lecturers were significant (Spring 2004: $df = 1, F = 11.89, p = .001$; Fall 2004: $df = 1, F = 5.77, p = .02$), as were the differences in the gains (Spring 2004: $df = 1, F = 11.89, p = .001$; Fall 2004: $df = 1, F = 5.77, p = .02$) and the standardized gains (Spring 2004: $df = 1, F = 8.07, p = .005$; Fall 2004: $df = 1, F = 3.80, p = .05$), with the average posttest score, gain, and standardized gain being higher for Lecturer 1 than in for the other lecturers.

Although these differences between lecturers obtained, they do not provide a *direct* test of whether students who (regardless of lecture) constructed correct argument diagrams have better skills. Although the students of Lecturer 1 were the only students to be explicitly taught how to construct argument diagrams, a substantial number of students of other lecturers constructed correct argument diagrams on their posttests. In addition, a substantial number of the students of Lecturer 1 constructed incorrect argument diagrams on their posttests. Thus, to test whether it was actually the construction of these diagrams that contributed to the difference in scores of the students of Lecturer 1, or whether is was the additional teaching methods of the Lecturer 1, we introduced a new variable into our model.

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.

The second hypothesis implies that the number of correct argument diagrams a student constructed on the posttest was correlated to the student's posttest score, gain and standardized gain. 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. 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 4) on the posttest. The results for Spring 2004 are given in Table 7.

TABLE 7
Spring 2004: Mean fractional score (standard deviation) for the pretest and the posttest, mean gain (standard deviation), and mean standardized gain (standard deviation)

	Pre	Post	Gain	StGain
No Correct	0.56 (0.02)	0.74 (0.02)	0.18 (0.02)	0.39 (0.03)
Few Correct	0.57 (0.02)	0.75 (0.02)	0.17 (0.02)	0.37 (0.04)
Many Correct	0.66 (0.02)	0.88 (0.01)	0.22 (0.02)	0.56 (0.06)

Similar data and results 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, 4, or 5) on the posttest. The results for Fall 2004 are given in Table 8.

TABLE 8
Fall 2004: Mean fractional score (standard deviation) for the pretest and the posttest, mean gain (standard deviation), and mean standardized gain (standard deviation)

	Pre	Post	Gain	StGain
No Correct	0.41 (0.02)	0.59 (0.03)	0.18 (0.02)	0.30 (0.04)
Few Correct	0.42 (0.03)	0.61 (0.02)	0.19 (0.03)	0.27 (0.04)
Many Correct	0.59 (0.04)	0.82 (0.02)	0.23 (0.03)	0.50 (0.06)

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 again indicates that the differences in the pretest scores was significant for predicting the posttest scores (Spring 2004: $df = 1, F = 23.67, p < .001$; Fall 2004: $df = 1, F = 41.87, p < .001$), 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: $df = 1, F = 28.66, p < .001$).

In addition, this analysis indicates that in each semester, even accounting for differences in pretest score, the differences in the posttest scores 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 differences in the gains (Spring 2004:

$df = 1, F = 28.13, p < .001$; Fall 2004: $df = 1, F = 37.78, p < .001$) and the standardized gains (Spring 2004: $df = 1, F = 22.27, p < .001$; Fall 2004: $df = 1, F = 34.14, p < .001$), with the average posttest score, gain, and standardized gain being higher for those who constructed many correct argument diagrams than for those who did not.

In both semesters the average posttest score was approximately 0.7, and the average gain and standardized gain from pretest to posttest was approximately 0.2 and 0.5, respectively. Using these numbers we can see very clearly the differences between the students who constructed many argument diagrams and those who constructed no or few correct argument diagrams on the posttest by comparing the frequency of students in each group who score below average on each measure to the frequency of students in each group who score above average on each measure. The comparisons of these frequencies are given in Figures 2-7.

These results show that the students who mastered the use of argument diagrams—those who constructed 3 or 4 correct argument diagrams for Spring 2004, or 3,4, or 5 correct argument diagrams for Fall 2004—had the highest posttest scores, gained the most from pretest to posttest, and gained the most as a fraction of the gain that was possible. 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 sections 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.

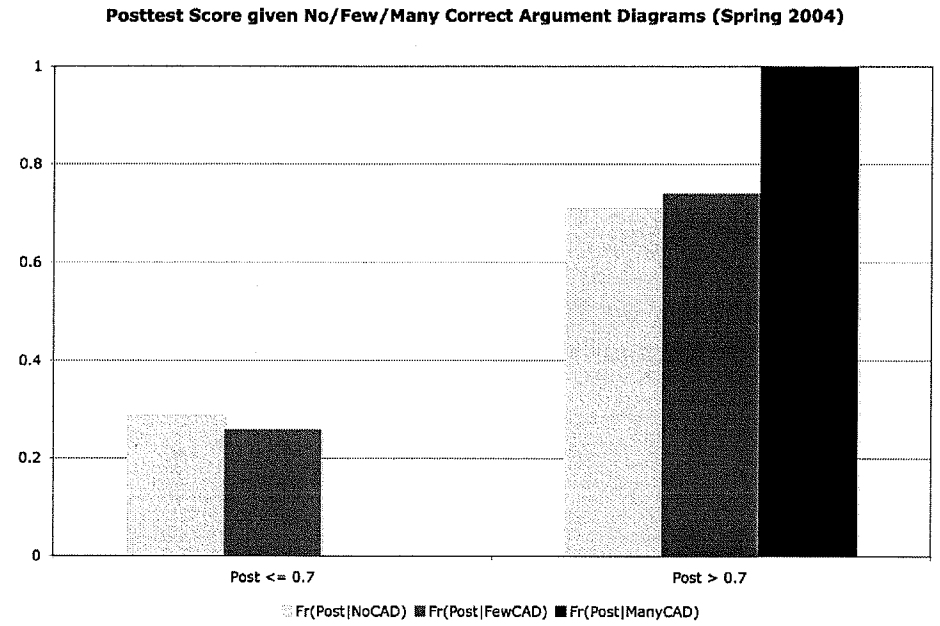

FIGURE 2 Histograms comparing the frequency of students (Spring 2004) who scored less than or equal to 0.7, and greater than 0.7 on the posttest given that they constructed no correct argument diagrams on the posttest to the frequency of students who scored less than or equal to 0.7, and greater than 0.7 on the posttest given that they constructed few (1 or 2) correct argument diagrams on the posttest and to the frequency of students who scored less than or equal to 0.7, and greater than 0.7 on the posttest given that they constructed many (3 or 4) correct argument diagrams on the posttest.

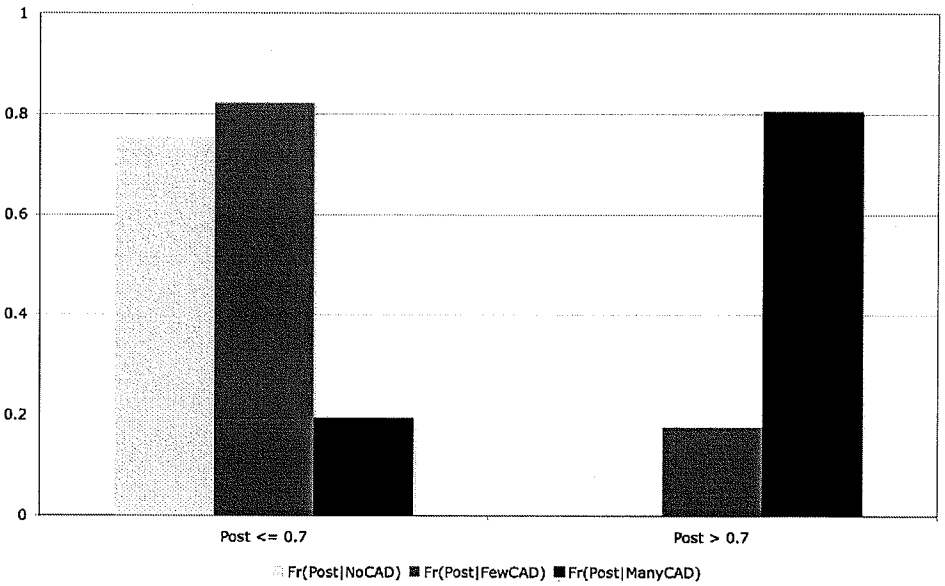

FIGURE 3 Histograms comparing the frequency of students (Fall 2004) who scored less than or equal to 0.7, and greater than 0.7 on the posttest given that they constructed no correct argument diagrams on the posttest to the frequency of students who scored less than or equal to 0.7, and greater than 0.7 on the posttest given that they constructed few (1 or 2) correct argument diagrams on the posttest and to the frequency of students who scored less than or equal to 0.7, and greater than 0.7 on the posttest given that they constructed many (3, 4 or 5) correct argument diagrams on the posttest.

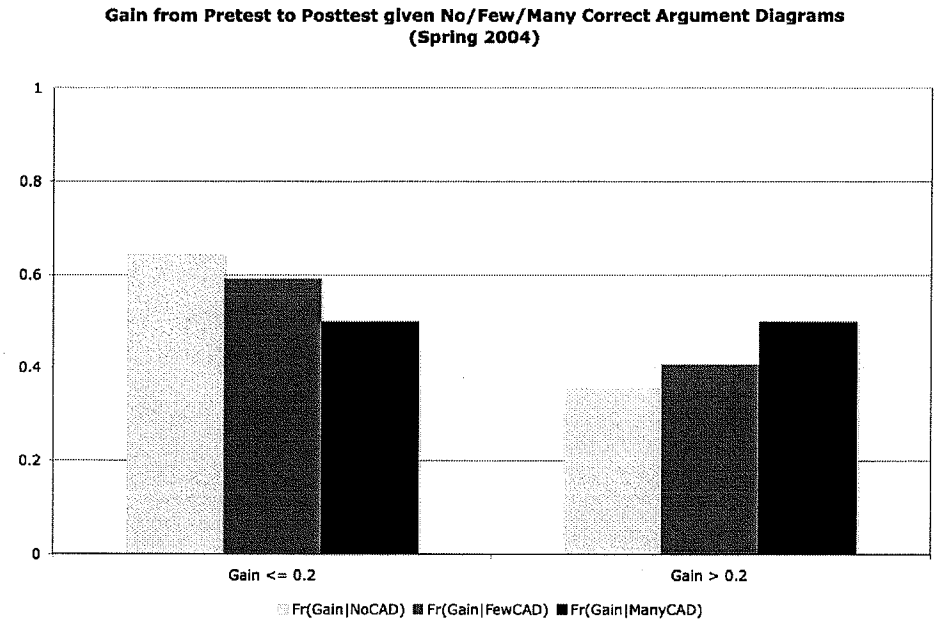

FIGURE 4 Histograms comparing the frequency of students (Spring 2004) who gained less than or equal to 0.2, and greater than 0.2 from pretest to posttest given that they constructed no correct argument diagrams on the posttest to the frequency of students who gained less than or equal to 0.2, and greater than 0.2 from pretest to posttest given that they constructed few (1 or 2) correct argument diagrams on the posttest and to the frequency of students who gained less than or equal to 0.2, and greater than 0.2 from pretest to posttest given that they constructed many (3 or 4) correct argument diagrams on the posttest.

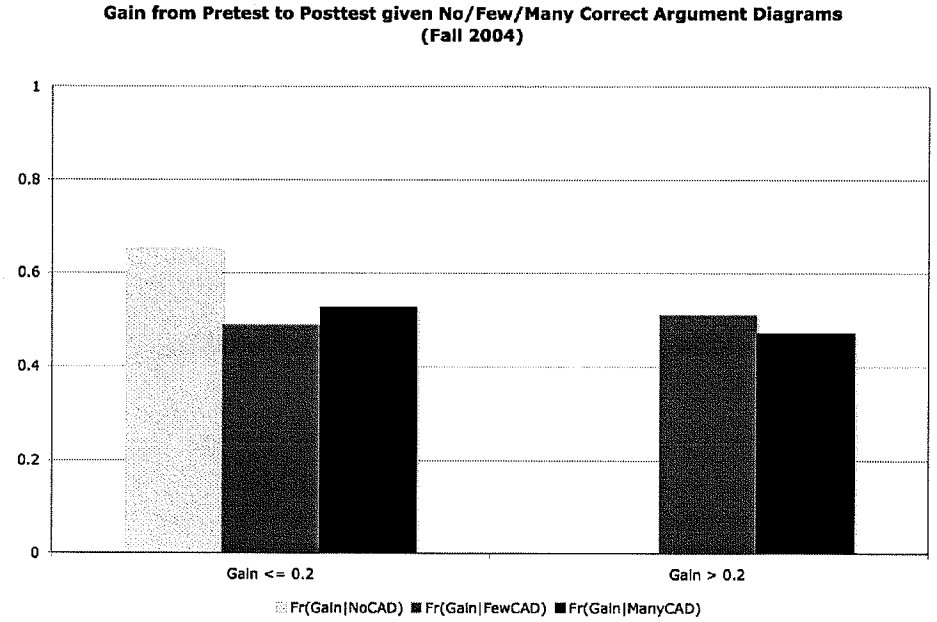

FIGURE 5 Histograms comparing the frequency of students (Fall 2004) who gained less than or equal to 0.2, and greater than 0.2 from pretest to posttest given that they constructed no correct argument diagrams on the posttest to the frequency of students who gained less than or equal to 0.2, and greater than 0.2 from pretest to posttest given that they constructed few (1 or 2) correct argument diagrams on the posttest and to the frequency of students who gained less than or equal to 0.2, and greater than 0.2 from pretest to posttest given that they constructed many (3, 4 or 5) correct argument diagrams on the posttest.

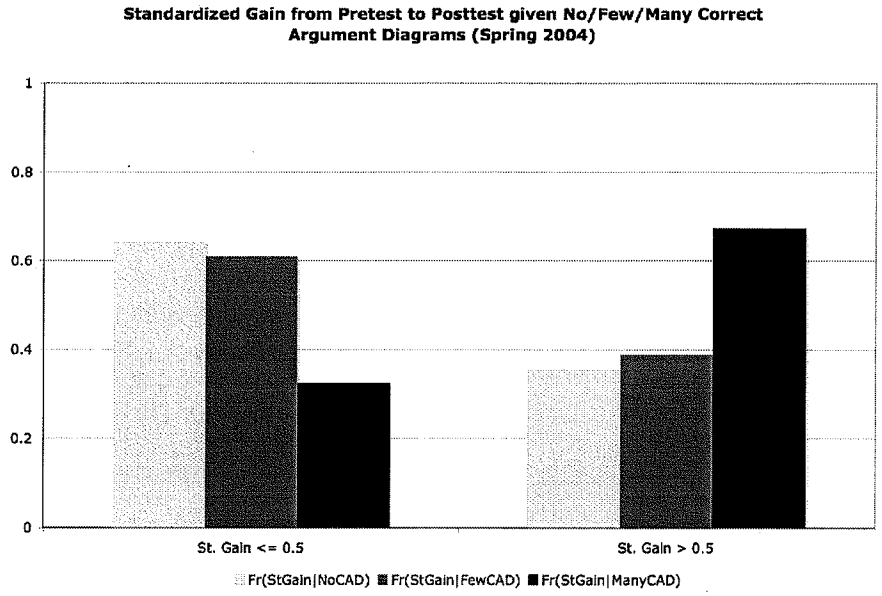

FIGURE 6 Histograms comparing the frequency of students (Spring 2004) who had a standardized gain less than or equal to 0.5, and greater than 0.5 from pretest to posttest given that they constructed no correct argument diagrams on the posttest to the frequency of students who had a standardized gain less than or equal to 0.5, and greater than 0.5 from pretest to posttest given that they constructed few (1 or 2) correct argument diagrams on the posttest and to the frequency of students who had a standardized gain less than or equal to 0.5, and greater than 0.5 from pretest to posttest given that they constructed many (3 or 4) correct argument diagrams on the posttest.

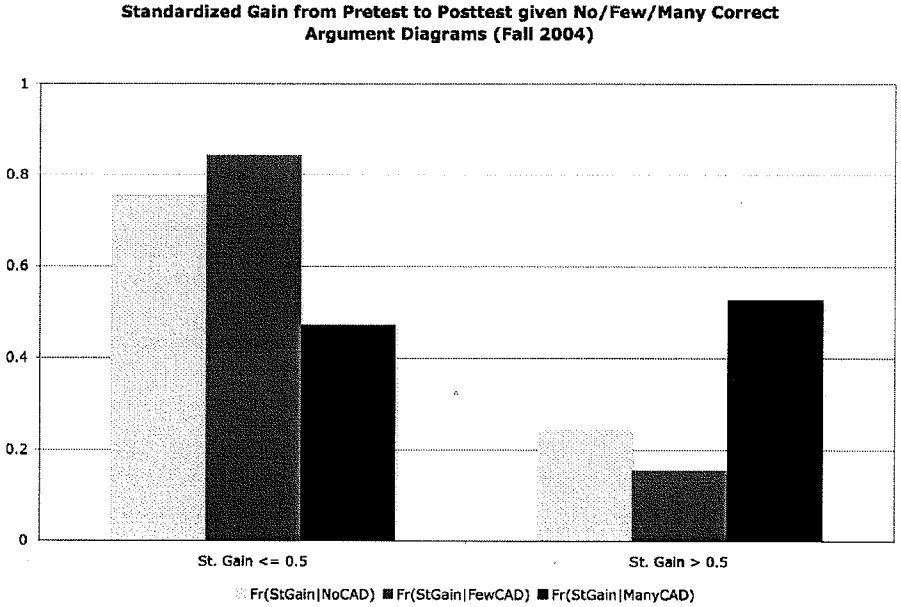

FIGURE 7 Histograms comparing the frequency of students (Fall 2004) who had a standardized gain less than or equal to 0.5, and greater than 0.5 from pretest to posttest given that they constructed no correct argument diagrams on the posttest to the frequency of students who had a standardized gain less than or equal to 0.5, and greater than 0.5 from pretest to posttest given that they constructed few (1 or 2) correct argument diagrams on the posttest and to the frequency of students who had a standardized gain less than or equal to 0.5, and greater than 0.5 from pretest to posttest given that they constructed many (3, 4 or 5) correct argument diagrams on the posttest.

D. 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 8 and 9.

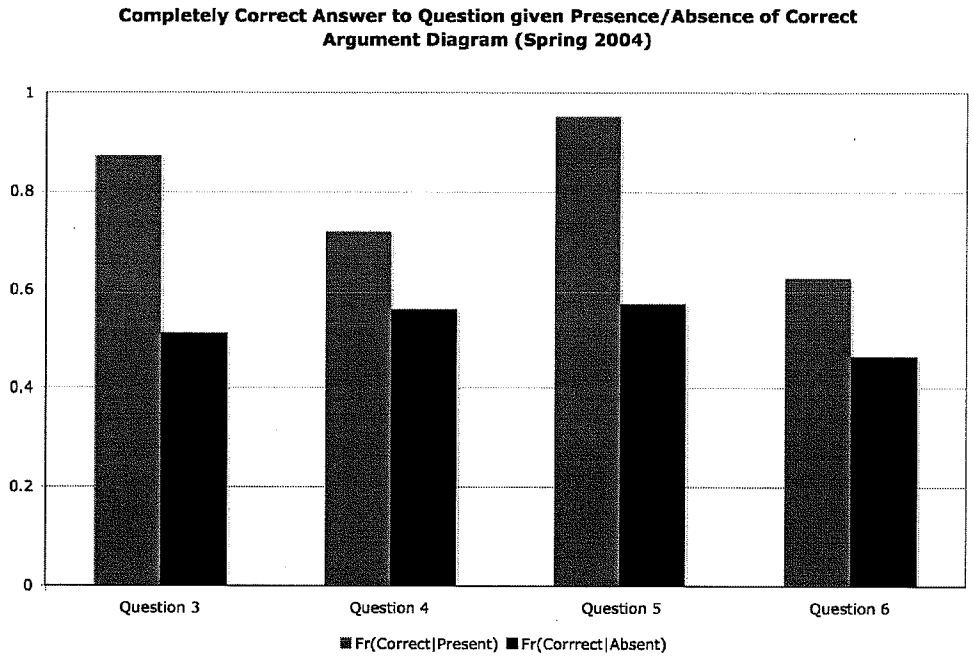

FIGURE 8 Histograms comparing the frequency of students (Spring 2004) who answered all parts of each question correctly given that they constructed the correct argument diagram for that question to the frequency of students who answered all parts of each question correctly given that they did not construct the correct argument diagram for that question.

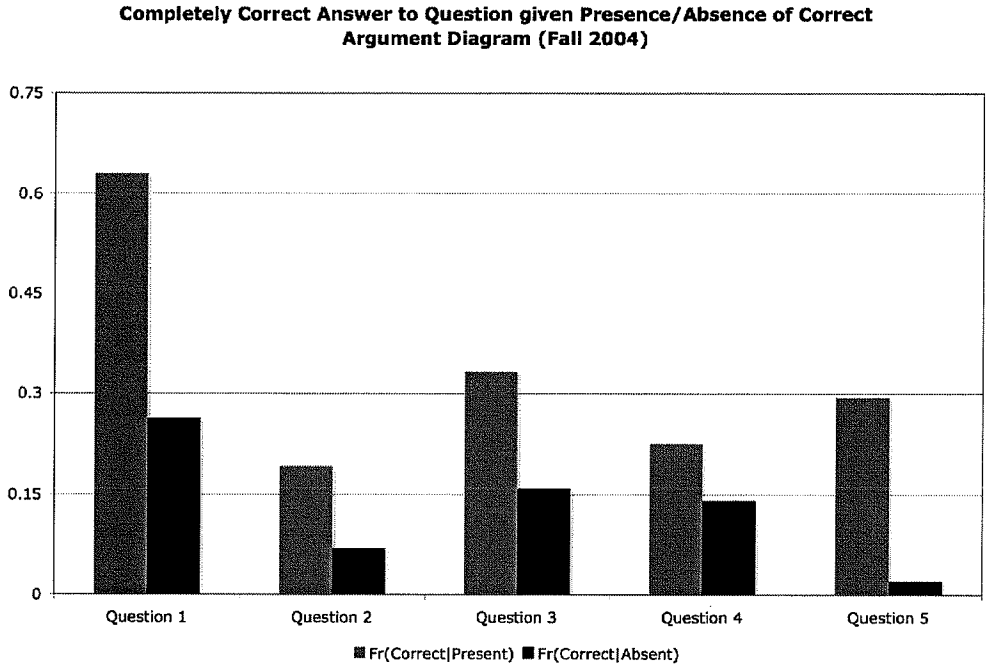

FIGURE 9 Histograms comparing the frequency of students (Fall 2004) who answered all parts of each question correctly given that they constructed the correct argument diagram for that question to the frequency of students who answered all parts of each question correctly given that they did not construct the correct argument diagram for that question.

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.

E. Prediction of Posttest Score, 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 lecture, it is possible that the use of argument diagrams was correlated with instructor's teaching ability, the student's year in school, etc.

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 three linear regressions—one for the posttest fractional score, a second for the gain, and a third 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 posttest fractional score, 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

Ignoring the variables that were not significant for either semester, we ran the regressions again. The two regression equations for each predicted variable for each semester are as follows:

Spring 2004 Posttest

Post	= 0.534	+ 0.306 Pre	+ 0.122 Lecturer1	+ 0.071 Lecturer2	+ 0.080 Lecturer3	
	(0.036)	(0.062)	(0.025)	(0.024)	(0.024)	
	$p < .001$	$p < .001$	**$p < .001$**	**$p = .004$**	**$p = .001$**	
Post	= 0.548	+ 0.244 Pre	+ 0.052 Lecturer1	+ 0.076 Lecturer2	+ 0.040 Lecturer3	+ 0.034 PostCAD
	(0.035)	(0.062)	(0.031)	(0.023)	(0.026)	(0.010)
	$p < .001$	$p < .001$	**$p = .096$**	**$p = .001$**	**$p = .131$**	**$p = .001$**

Fall 2004 Posttest

Post	= 0.505	+ 0.343 Pre	+ 0.082 Lecturer1	+ 0.023 Lecturer2	– 0.114 Lecturer5	
	(0.031)	(0.067)	(0.039)	(0.030)	(0.032)	
	$p < .001$	$p < .001$	**$p = .035$**	**$p = .468$**	**$p < .001$**	
Post	= 0.444	+ 0.212 Pre	+ 0.074 Lecturer1	+ 0.112 Lecturer2	– 0.026 Lecturer5	+ 0.053 PostCAD
	(0.030)	(0.064)	(0.035)	(0.031)	(0.032)	(0.009)
	$p <$ .001	$p = .001$	**$p = .034$**	**$p < .001$**	**$p = .410$**	**$p < .001$**

Spring 2004 Gain

Gain	= 0.534	– 0.694 Pre	+ 0.122 Lecturer1	+ 0.071 Lecturer2	+ 0.080 Lecturer3	
	(0.036)	(0.062)	(0.025)	(0.024)	(0.024)	
	$p < .001$	$p < .001$	**$p < .001$**	**$p = .004$**	**$p = .001$**	
Gain	= 0.548	– 0.756 Pre	+ 0.052 Lecturer1	+ 0.076 Lecturer2	+ 0.040 Lecturer3	+ 0.034 PostCAD
	(0.035)	(0.062)	(0.031)	(0.023)	(0.026)	(0.010)
	$p < .001$	$p < .001$	**$p = .096$**	**$p = .001$**	**$p = .131$**	**$p = .001$**

Fall 2004 Gain

Gain	= 0.505	– 0.657 Pre	+ 0.082 Lecturer1	+ 0.023 Lecturer2	– 0.114 Lecturer5	
	(0.031)	(0.067)	(0.039)	(0.030)	(0.032)	
	$p < .001$	$p < .001$	**$p = .035$**	**$p = .468$**	**$p < .001$**	
Gain	= 0.444	– 0.788 Pre	+ 0.074 Lecturer1	+ 0.112 Lecturer2	– 0.026 Lecturer5	+ 0.053 PostCAD
	(0.030)	(0.064)	(0.035)	(0.031)	(0.032)	(0.009)
	$p < .001$	$p = .005$	**$p = .034$**	**$p < .001$**	**$p = .410$**	**$p < .001$**

Spring 2004 Standardized Gain

StGain	= 0.818	– 0.948 Pre	+ 0.305 Lecturer1	+ 0.199 Lecturer2	+ 0.209 Lecturer3
	(0.103)	(0.176)	(0.069)	(0.069)	(0.069)
	$p < .001$	$p < .001$	**$p < .001$**	**$p = .004$**	**$p = .003$**

StGain	= 0.851	– 1.096 Pre	+ 0.136 Lecturer1	+ 0.211 Lecturer2	+ 0.112 Lecturer3	+ 0.083 PostCAD
	(0.101)	(0.179)	(0.090)	(0.067)	(0.075)	(0.029)
	$p < .001$	$p < .001$	**$p = .132$**	**$p = .002$**	**$p = .138$**	**$p = .005$**
Fall 2004 Standardized Gain						
StGain	= 0.623	– 0.659 Pre	+ 0.169 Lecturer1	+ 0.080 Lecturer2	– 0.188 Lecturer5	
	(0.068)	(0.069)	(0.084)	(0.065)	(0.069)	
	$p < .001$	$p < .001$	**$p = .048$**	**$p = .223$**	**$p = .007$**	
StGain	= 0.494	– 0.951 Pre	+ 0.150 Lecturer1	+ 0.281 Lecturer2	– 0.009 Lecturer5	+ 0.118 PostCAD
	(0.065)	(0.139)	(0.075)	(0.067)	(0.069)	(0.020)
	$p < .001$	$p < .001$	**$p = .046$**	**$p < .001$**	**$p = .902$**	**$p < .001$**

These results show that in each set of regressions a student's pretest score was a highly significant predictor of the posttest score, gain, and standardized gain. In each case the coefficient of the pretest was positive when predicting the posttest, as expected; if all the students' scores generally improve from the pretest to the posttest, we expect the students who scored higher on the pretest to score higher on the posttest.

In addition, in each case, the coefficient of the pretest was negative when predicting gain and standardized gain. In fact, since the score on the pretest is a part of the value of the gain and standardized gain, it is interesting that the coefficient for pretest was significant at all. However, a regression run on a model that predicts gain and standardized gain based on all the above variables *except* the pretest shows that none of the variables are significant. We believe that this can be explained by the fact that scores on the pretest were not evenly distributed throughout the lectures, as we can see from Tables 5 and 6. The correlations between which lecturer a student had and his or her score on the pretest are given in Tables 11 and 12.

TABLE 11
Spring 2004: Pearson correlation between Pre and Lecturer 1, Lecturer 2, Lecturer 3, and Lecturer 4

	Lecturer 1	Lecturer 2	Lecturer 3	Lecturer 4
Pre	0.203*	0.133	-0.052	-0.281**

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

TABLE 12
Fall 2004: Pearson correlation between Pre and Lecturer 1, Lecturer 2, Lecturer 4, Lecturer 5 and Lecturer 6

	Lecturer 1	Lecturer 2	Lecturer 4	Lecturer 5	Lecturer 6
Pre	0.512***	0.131	-0.404***	-0.272**	0.015

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

So, a plausible explanation for the negative coefficient when predicting gain is that the students who scored the lowest on the pretest gained the most—and this is to be expected at least because there is more room for them to improve. In addition, a plausible explanation for the negative coefficient when predicting standardized gain is that, since the grade a student received on the posttest counted as a part of his or her grade in the course, the students who scored the lowest on the pretest had more incentive to improve, and thus, as a percentage of what they could have gained, gained more than the students who scored highest on the pretest. Thus, since we are also

concluding that there is a correlation between the lecturer the student had and the score on the posttest, gain, and standardized gain (see below), there are many contributing factors to a student's gain—the score on the pretest being one—which may be roughly offset if all the relevant variables are not examined.

From the results of the regression analysis we can also see that in both semesters, before we introduced the variable PostCAD, the coefficient for Pre was significantly positive for predicting posttest score and significantly negative for predicting gain and standardized gain. In addition, the coefficient for Lecturer 1 was significantly positive for predicting a student's posttest score and gain, and standardized gain. In addition, the coefficients for Lecturer 3 are significantly positive, while the coefficients for Lecturer 5 are significantly negative for predicting a student's posttest score and 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 posttest score and gain, and standardized gain.

From the results of the regression analysis we can see that in both semesters, after we introduce the variable PostCAD, the coefficient for Pre remains significant, but has a reduced value for each measure. In addition, in the Spring of 2004 when including the variable PostCAD, the variables Lecturer 1 and Lecturer 3 are no longer significant as predictors of posttest score, 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 posttest score, 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 posttest score, 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 Lecture 2 was still significantly positive for predicting a student's posttest score, 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 posttest score, 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 posttest score, 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—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 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.

These data support two simple causal pictures. Since, in both semesters, a student's pretest score is a significant predictor of his or her posttest score, gain and standardized gain, no matter which other variables are involved in the regression, we conjecture that a student's pretest score has a direct positive causal influence on the student's posttest score, and a negative causal influence on the student's gain and standardized gain. However, the coefficient for the variable Pre changes slightly when we add the variable PostCAD as a regressor, indicating that Pre is correlated with PostCAD (see Tables 13 and 14). We conjecture that this is because a student's score on the pretest is significantly correlated with the lecture in which he or she was enrolled (see Table 4), and the lecture a student was enrolled in is significantly correlated to the number of correct argument diagrams the student constructed on the posttest. Thus we conjecture that there is an unknown common cause of the variables Pre and Lecture 1.

TABLE 13
Spring 2004: Pearson correlation between PostCAD and Pre, Lecturer 1, Lecturer 2, Lecturer 3, and Lecturer 4

	Pre	Lecturer 1	Lecturer 2	Lecturer 3	Lecturer 4
PostCAD	0.318***	0.636***	-0.413***	0.174*	-0.384***

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

TABLE 14
Fall 2004: Pearson correlation between PostCAD and Pre, Lecturer 1, Lecturer 2, Lecturer 4, Lecturer 5 and Lecturer 6

	Pre	Lecturer 1	Lecturer 2	Lecturer 4	Lecturer 5	Lecturer 6
PostCAD	0.412***	0.465***	-0.334***	0.148	-0.385***	0.181*

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

In addition, as noted above, in the Spring of 2004 the coefficient for Lecture 1 is significantly positive for predicting a student's posttest score, gain and standardized gain before the variable PostCAD is introduced as a regressor, but insignificant afterwards. In addition, when introduced, the coefficient for PostCAD is significantly positive for predicting a student's posttest score, gain and standardized gain. On the other hand, in the Fall of 2004, the coefficient for Lecture 1 is

significantly positive for predicting a student's posttest score, gain and standardized gain both before and after the variable PostCAD is introduced as a regressor. But, as in the Spring of 2004, when introduced, the coefficient for PostCAD is significantly positive for predicting a student's posttest score, gain and standardized gain. Thus we conjecture that the number of correct argument diagrams a student constructed on the posttest has a direct positive causal influence on these measures, and, since being taught by Lecturer 1 was significantly correlated with constructing correct argument diagrams (see Tables 13 and 14) that in the Spring of 2004 whether a student was taught by Lecturer 1 has a positive causal influence only on whether he or she constructed correct argument diagrams on the posttest (relative to students enrolled in the other lectures); that is, whether a student was taught by Lecturer 1 does not have a direct causal influence on his or her posttest score, gain or standardized gain. However, in the Fall of 2004, whether a student was taught by Lecturer 1 has a positive causal influence not only on whether he or she constructed correct argument diagrams on the posttest (relative to students enrolled in the other lectures), but also directly on his or her posttest score, gain or standardized gain.

The situation seems slightly different for the Lecturer 2, however. In the Spring of 2004, the coefficient for Lecturer 2 is significantly positive for predicting a student's posttest score, gain and standardized gain both before and after the variable PostCAD is introduced as a regressor. Since the students of Lecturer 2 were significantly less likely to construct correct argument diagrams (see Table 13), we conjuecture that whether a student was taught by Lecturer 2 has a direct negative influence on whether a student constructed correct argument diagrams, and that whether a student was taught by Lecturer 2 has a direct positive causal influence on his or her posttest score, gain and standardized gain—a positive influence that greatly outweighed the effects of the negative influence on constructing correct argument diagrams. In the Fall of 2004, on the other hand, the coefficient for Lecturer 2 is significantly positive for predicting a student's posttest score, gain and standardized gain only after the variable PostCAD is introduced as a regressor. Since the students of Lecturer 2 were again significantly less likely to construct correct argument diagrams (see Table 14), we conjuecture that whether a student was taught by Lecturer 2 has a direct negative influence on whether a student constructed correct argument diagrams, and that whether a student was taught by Lecturer 2 has a direct positive causal influence on his or her posttest score, gain and standardized gain—a positive influence that just made up for the effects of the negative influence on constructing correct argument diagrams.

Additionally, in the Spring of 2004, the coefficient for Lecturer 3 was significantly positive for predicting posttest score, gain and standardized gain only before the variable PostCAD was introduced as a regressor. Since being taught by Lecturer 3 was significantly positively correlated with constructing correct argument diagrams (even though Lecturer 3 did not explicitly teach argument diagramming, see Table 13), we conjecture that whether a student was taught by Lecturer 3 has a positive causal influence only on whether he or she constructed correct argument diagrams on the posttest (relative to students enrolled in the other lectures); that is, whether a student was taught by Lecturer 3 does not have a direct causal influence on his or her posttest score, gain or standardized gain.

And similarly in the Fall of 2004, the coefficient for Lecturer 5 was significantly negative for predicting posttest score, gain and standardized gain only before the variable PostCAD was introduced as a regressor. Since being taught by Lecturer 5 was significantly negatively

correlated with constructing correct argument diagrams (see Table 14), we conjecture that whether a student was taught by Lecturer 5 has a negative causal influence only on whether he or she constructed correct argument diagrams on the posttest (relative to students enrolled in the other lectures); that is, whether a student was taught by Lecturer 5 does not have a direct causal influence on his or her posttest score, gain or standardized gain.

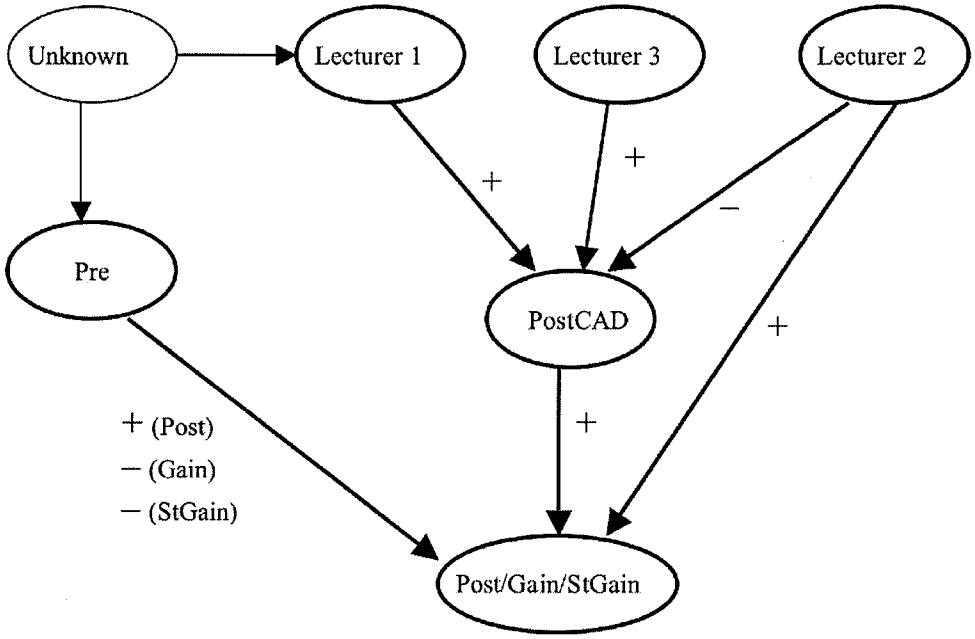

FIGURE 10 A diagram representing a plausible picture of the causal links between the variables that are significant predictors of posttest score, gain and standardized gain for Spring 2004.

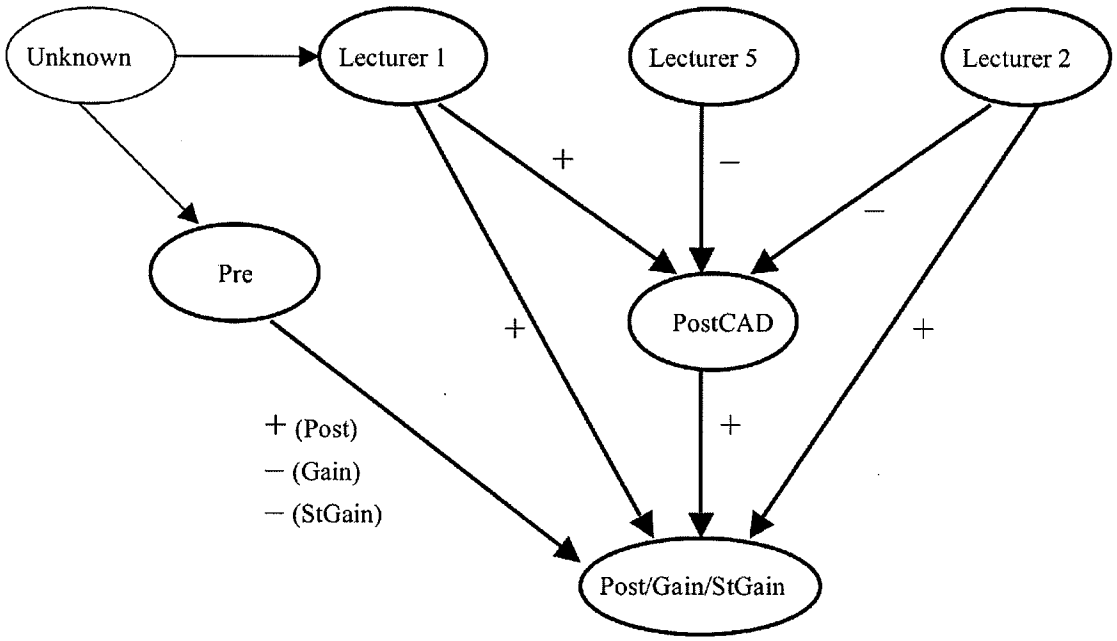

Figure 11 A diagram representing a plausible picture of the causal links between the variables that are significant predictors of posttest score, gain and standardized gain for Fall 2004.

The causal pictures described above can be represented as a causal diagram, shown in Figures 10 and 11.

A stronger version of our second hypothesis, then, is confirmed by these results: constructing correct argument diagrams not only positively contributes to the improvement of argument analysis, but can also over-shadow differences in instruction and personal history.

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.

Educational Importance

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.

Perhaps, though, both the colleges and employers believe that the ability to reason well is the kind of skill that is taught not intensively in any one course, but rather across the curriculum, in a way that would ensure that students acquired these skills no matter what major they chose. The research seems to show, however, that this is not the case; on tests of general critical thinking skills, students average a gain of less than one standard deviation during their entire time in college, while most of this gain comes just in the first year.

In fact, these are among the reasons we give to prospective majors for joining the philosophy department. We can cite statistics about which majors generally do better on the LSAT and GRE; but what we have not been able to do in the past is show evidence that our classes improve critical thinking skills.

What this study shows is that students do 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. 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.

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.

Descartes, R. (1641). *Meditations on First Philosophy*. Edited and translated by E. Haldane and G.R.T. Ross in *The Philosophical Works of Descartes, Vol. 1*. Cambridge University Press, 1969.

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

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.

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.

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* (pp. 177-189). Philadelphia: The Franklin Institute Press.

Plato. (1976). *Meno*. Translated by G.M.A. Grube. Indianapolis: Hackett.

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. (2204) 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* (pp. 97-115). New York: Springer.

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

[1] 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