

Using Argument Diagrams to Improve  
Critical Thinking Skills  
in Introductory Philosophy

Maralee Harrell

Carnegie Mellon University  
Department of Philosophy  
135 Baker Hall  
Pittsburgh, PA, 15213  
mharrell@cmu.edu  
(412) 268-8152

**Abstract**

In an experiment involving 139 students in an introductory philosophy course we tested whether students were improving their ability to think critically about arguments and whether using argument diagramming as an analysis aid contributed to this improvement. We determined that the students did develop this skill over the course of the semester. We also determined that the students in one section of the course gained significantly more than the students in the other sections, and that this was due almost entirely to their ability to use argument diagrams. We conclude that learning how to construct argument diagrams significantly improves a student's ability to analyze, comprehend, and evaluate arguments.

## Using Argument Diagrams to Improve Critical Thinking Skills in Introductory Philosophy

In the introductory philosophy class at Carnegie Mellon University (*80-100 What Philosophy Is*) one important learning goal is the development of general critical thinking skills. Even though there are a few generally accepted measures of these skills (e.g. the California Critical Thinking Skills Test and the Watson Glaser Critical Thinking Appraisal, but see also Halpern, 1989 and Paul, Binker, Jensen, & Kreklau, 1990), there is surprisingly little research on the sophistication of, or on effective methods for improving, the critical thinking skills of college students. 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 actually improve these skills (Annis & Annis, 1979; Pascarella, 1989; Stenning, Cox, & Oberlander, 1995).

Critical thinking involves the ability to analyze, understand, and evaluate an argument. Our first hypothesis is that students improved on these tasks after taking the introductory philosophy course. However, we wanted to determine not only whether they improved, but how much improvement could be attributed to alternative teaching methods.

One candidate method is the use of argument diagrams as an aid to overall argument comprehension, since we believe that they significantly facilitate understanding, analysis, and evaluation. An argument is a series of statements in which one is the conclusion, and the others are premises supporting this conclusion; and an argument diagram is a visual representation of these statements and the inferential connections between them.

For example, at the end of *Meno*, Plato (1976) argues through the character of Socrates that virtue is a gift from the gods (89d-100b). While the English translations of Plato's works are among the more readable philosophical texts, it is still the case 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 Plato uses to support his conclusion, and visually represent the connections between these statements (as shown in Figure 1), the structure of the argument is immediately clear, as are the places where we may critique or applaud it.

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, et al. 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. However, to our knowledge there has been no research done to determine whether it is the mere ability to construct argument diagrams, or the aid of a computer platform and tutor (or possibly both) that is the crucial factor.

Our second hypothesis is that the crucial factor in the improvement of critical thinking skills is the ability to construct argument diagrams. This hypothesis posits that students who construct correct diagrams during argument analysis tasks should perform better on these tasks than students who do not.

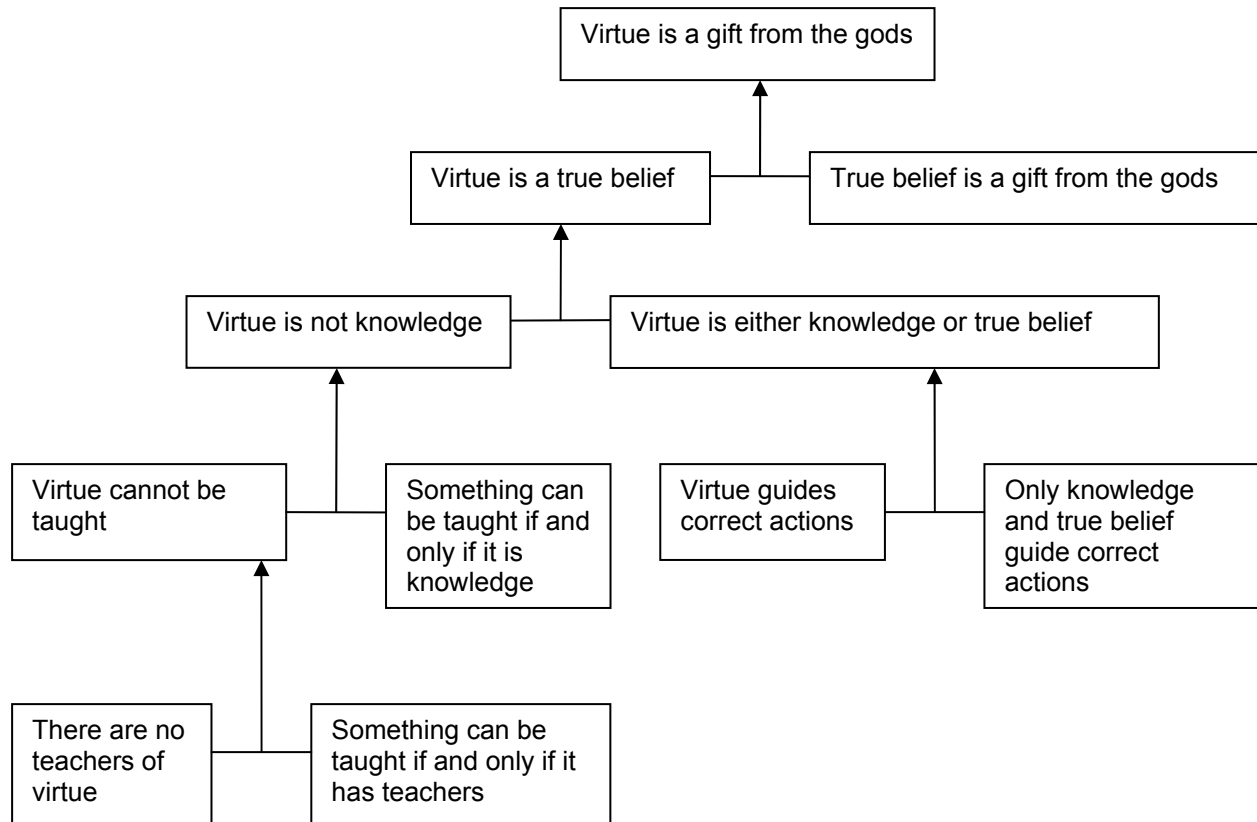


FIGURE 1 An argument diagram representing one of the arguments in Plato's *Meno*.

We typically teach several sections of Carnegie Mellon University's introduction to philosophy course (*80-100 What Philosophy Is*) each semester, with a different instructor for each section. While the general curriculum of the course is set, each instructor is given a great deal of freedom in executing this curriculum. For example, each section is a topics based course in which epistemology, metaphysics, and ethics are introduced with both historical and contemporary primary-source readings. Each instructor, however, chooses a text, the order of the topics, and the assignments for his or her section. The students who take this course are a mix of classes and majors from all over the University.

In the Spring of 2004, students in Section 1 were explicitly taught how to construct argument diagrams to represent a selection of text. In contrast, students in Sections 2, 3, and 4 were not explicitly taught the use of argument diagrams, but rather—if they were taught to analyze arguments at all—were taught to use more traditional kinds of representations (e.g. lists of statements).

In this study, we test the first hypothesis by comparing the pretest and posttest scores of all the students in 80-100 in the Spring semester of 2004. We test the second hypothesis in three ways: (1) by comparing the pretest and posttest scores of students in Section 1 to students in Sections 2, 3, and 4, (2) by comparing the pretest and posttest scores of students who constructed correct argument diagrams on the posttest to those students who did not, and (3) by comparing total scores on individual questions on the posttest of students who constructed the correct argument diagrams for that question to those students who did not.

## Method

### *Participants*

139 students (46 women, 93 men) in each of 4 sections of introductory philosophy (*80-100 What Philosophy Is*) at Carnegie Mellon University in the Spring of 2004 were studied. Each section of the course had a different instructor and teaching assistant, and the students chose their section. There were 35 students (13 women, 22 men) in Section 1, 37 students (18 women, 19 men) in Section 2, 32 students (10 women, 22 men) in Section 3, and 35 students (5 women, 30 men) in Section 4. The students in Section 1 were taught the use of argument diagrams to analyze the arguments in the course reading, while the students in the other three sections were taught more traditional methods of analyzing arguments.

### *Materials*

Prior to the semester, the four instructors of 80-100 in the Spring of 2004 met to determine the learning goals of this course, and designed an exam to test the students on relevant skills. 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 each question on the pretest, there was a structurally (nearly) identical question with different content on the posttest. 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) 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.

### *Procedure*

Each of the four sections of 80-100 was a Monday/Wednesday/Friday class. The pretest was given to all students during the second day of class. The students in sections 2 and 3 were given the posttest on the last day of classes, while the students in sections 1 and 4 were given the posttest as one part of their final exam, during exam week.

## Results and Discussion

### *Test Coding*

Pre- and posttests were paired by student—single-test students were excluded from the sample—so that there were 139 pairs of tests in the study. Tests which did not have pairs were used for coder-calibration, prior to the coding of the 139 pairs of tests.

Two graduate students independently coded all 278 tests (139 pairs). Each pre-/posttest 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. 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 pretest and one for the posttest) 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 the truth of the premises and/or the support of premises for the 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 (e.g.,  $P1 + P2/Conclusion (C)$ ), Other or blank.

To determine inter-coder reliability, the Percentage Agreement (PA), Cohen’s Kappa ( $\kappa$ ) and Krippendorff’s Alpha ( $\alpha$ ) were calculated for each test (given in Table 1).

TABLE 1  
Inter-coder Reliability: Percentage Agreement (PA),  
Cohen’s Kappa ( $\kappa$ ), and Krippendorff’s Alpha ( $\alpha$ ) for each test

	PA	$\kappa$	$\alpha$
<b>Pretest</b>	.85	.68	.68
<b>Posttest</b>	.85	.55	.54

The inter-coder reliability was fairly good, however, upon closer examination it was determined that one coder had systematically higher standards than the other coder 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). Specifically, on the 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 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.” In light of this, the codes from each coder 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 questions 3-6 were preliminarily set aside, while the addition of the codes received on questions 1 and 2, as well as parts (a), (b), (c), and (e) of questions 3-6 determined the raw score a student received on the test.

TABLE 2  
The variables and their descriptions recorded for each student

Variable Name	Variable Description
Pre	Fractional score on the pretest
Post	Fractional score on the posttest
A*	Averaged score (or code) on the pretest for question *
B*	Averaged score (or code) on the posttest for question *
Section	Enrolled section
Sex	Student’s sex
Honors	Enrollment in Honors course
Grade	Final Grade in the course
Year	Year in school

The primary variables of interest were the fractional pretest and posttest scores (the raw score converted into a percentage), 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 sex, and whether the student had taken the concurrent honors course associated with the introductory course. Table 2 gives summary descriptions of these variables.

*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 is given in Table 3.

TABLE 3  
Mean fractional score (standard deviation) for the pretest and the posttest,  
mean gain (standard deviation), and mean standardized gain (standard deviation)

	Pre	Post	Gain	GainSt.
<b>Whole Population</b>	0.59 (0.14)	0.78 (0.12)	0.19 (0.01)	0.43 (0.03)

The difference in the means of the pretest and posttest scores was significant (paired  $t$ -test;  $p < .001$ ). In addition, 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: overall the students did have significant gains and standardized gains from pretest to posttest.

*Comparison of Gains of Students by Section 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 in Section 1, we first tested this hypothesis by determining whether the average gain of the students in Section 1 was significantly different from the average gain of the students in each of the other sections. 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 section is 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	GainSt.
<b>Section 1</b>	0.64 (0.14)	0.85 (0.10)	0.21 (0.02)	0.51 (0.07)
<b>Section 2</b>	0.53 (0.16)	0.70 (0.14)	0.17 (0.03)	0.32 (0.05)
<b>Section 3</b>	0.58 (0.14)	0.79 (0.08)	0.21 (0.02)	0.48 (0.04)
<b>Section 4</b>	0.63 (0.10)	0.80 (0.09)	0.17 (0.02)	0.42 (0.05)

Since there was such variability in the scores on the pretest among the different sections, we ran an ANCOVA on the each of the variables Post, Gain, and GainSt, with the variable Pre used as the covariate. This analysis indicates that the differences in the pretest scores was significant for predicting the posttest scores ( $df = 1, F = 24.36, p < .001$ ), the gain ( $df = 1, F = 125.50, p < .001$ ), and the standardized gain ( $df = 1, F = 29.14, p < .001$ ). In addition, this analysis indicates that, even accounting for differences in pretest score, the differences in the posttest scores among the sections were significant ( $df = 3, F = 8.71, p < .001$ ), as were the differences in the gains ( $df = 3, F = 8.71, p < .001$ ) and the standardized gains ( $df = 3, F = 6.84, p < .001$ ).

This analysis shows that a student's section is a significant predictor of posttest score, gain, and standardized gain, but it does not tell us how they are different. The hypothesis is that the posttest score, gain and standardized gain for students in Section 1 is significantly higher than all the other sections. Thus, we did a planned comparison of the variables Post, Gain, and GainSt for Section 1 with the other sections 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 ( $df = 1, F = 32.28, p < .001$ ), the gain ( $df = 1, F = 107.37, p < .001$ ), and the standardized gain ( $df = 1, F = 21.42, p < .001$ ). In addition, this analysis indicates that, even accounting for differences in pretest score, the differences in the posttest scores between Section 1 and the other sections were significant ( $df = 1, F = 11.89, p = .001$ ), as were the differences in the gains ( $df = 1, F = 11.89, p = .001$ ) and the standardized gains ( $df = 1, F = 8.07, p = .005$ ), with the average posttest score, gain, and standardized gain being higher in Section 1 than in the other three sections.

Although these differences between sections (at least with standardized gain scores) obtained, they do not provide a direct test of whether students who (regardless of section) constructed correct argument diagrams have better skills. The explanation is that, although the students in Section 1 were the only students to be explicitly taught how to construct argument diagrams, a substantial number of students from other sections constructed correct argument diagrams on their posttests. In addition, a substantial number of the students in Section 1 constructed incorrect argument diagrams on their posttests. Thus, to test whether it was actually the construction of these diagrams that contributed to the higher scores of the students in Section 1, or whether it was the other teaching methods of the instructor for Section 1, we introduced a new variable into our model.

Recall that the type of answer given on part (d) of questions 3-6 was the data recorded from the test. 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 PostAD (value = 0, 1, 2, 3, 4).

The second hypothesis implies that the number of correct argument diagrams a student constructed on the posttest was correlated to the student's pretest score, posttest score, gain and standardized gain. Since there were very few students who constructed exactly 2 correct argument diagrams on the posttest, and still fewer who constructed exactly 4, we grouped the students by whether they had constructed No correct argument diagrams (PostAD = 0), Few correct argument diagrams (PostAD = 1 or 2), or Many correct argument diagrams (PostAD = 3 or 4) on the posttest. The results are given in Table 5.

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

	Pre	Post	Gain	GainSt.
<b>No Correct</b>	0.56 (0.16)	0.74 (0.12)	0.18 (0.02)	0.39 (0.03)
<b>Few Correct</b>	0.57 (0.13)	0.75 (0.12)	0.17 (0.02)	0.37 (0.04)
<b>Many Correct</b>	0.66 (0.13)	0.88 (0.06)	0.22 (0.02)	0.56 (0.06)

Again, since there was such variability in the scores on the pretest among the different sections, we ran an ANCOVA on each of the variables Post, Gain, and GainSt, with the variable Pre used as the covariate. This analysis indicates that the differences in the pretest scores was significant for predicting the posttest scores ( $df = 1, F = 24.68, p < .001$ ), the gain ( $df = 1, F = 132.81, p < .001$ ), and the standardized gain ( $df = 1, F = 30.97, p < .001$ ). This analysis also indicates that, even accounting for differences in pretest score, the differences among the students who constructed 0, Few or Many correct argument diagrams on the posttest are significant ( $df = 2, F = 14.66, p < .001$ ), as are the differences in gains ( $df = 2, F = 14.66, p < .001$ ), and standardized gains ( $df = 2, F = 11.78, p < .001$ ).

This analysis shows that whether a student constructed no, few or many correct argument diagrams is a significant predictor of posttest score, gain, and standardized gain, but it does not tell us how they are different. The hypothesis is that the posttest score, gain and standardized gain for students who constructed many diagrams is significantly different from both of the other groups. Thus, we did a planned comparison of the variables Post, Gain, and GainSt 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 ( $df = 1, F = 23.67, p < .001$ ), the gain ( $df = 1, F = 132.00, p < .001$ ), and the standardized gain ( $df = 1, F = 31.29, p < .001$ ). In addition, this analysis indicates that, 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 ( $df = 1, F = 28.13, p < .001$ ), as were the differences in the gains ( $df = 1, F = 28.13, p < .001$ ) and the standardized gains ( $df = 1, F = 22.27, p < .001$ ), 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.

These results show that the students who mastered the use of argument diagrams—those who constructed 3 or 4 correct argument diagrams—had the highest posttest scores 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.

### *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 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 questions 3-6 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 correlations between constructing the correct argument diagram and answering correctly all parts of each questions are given in Figure 2.

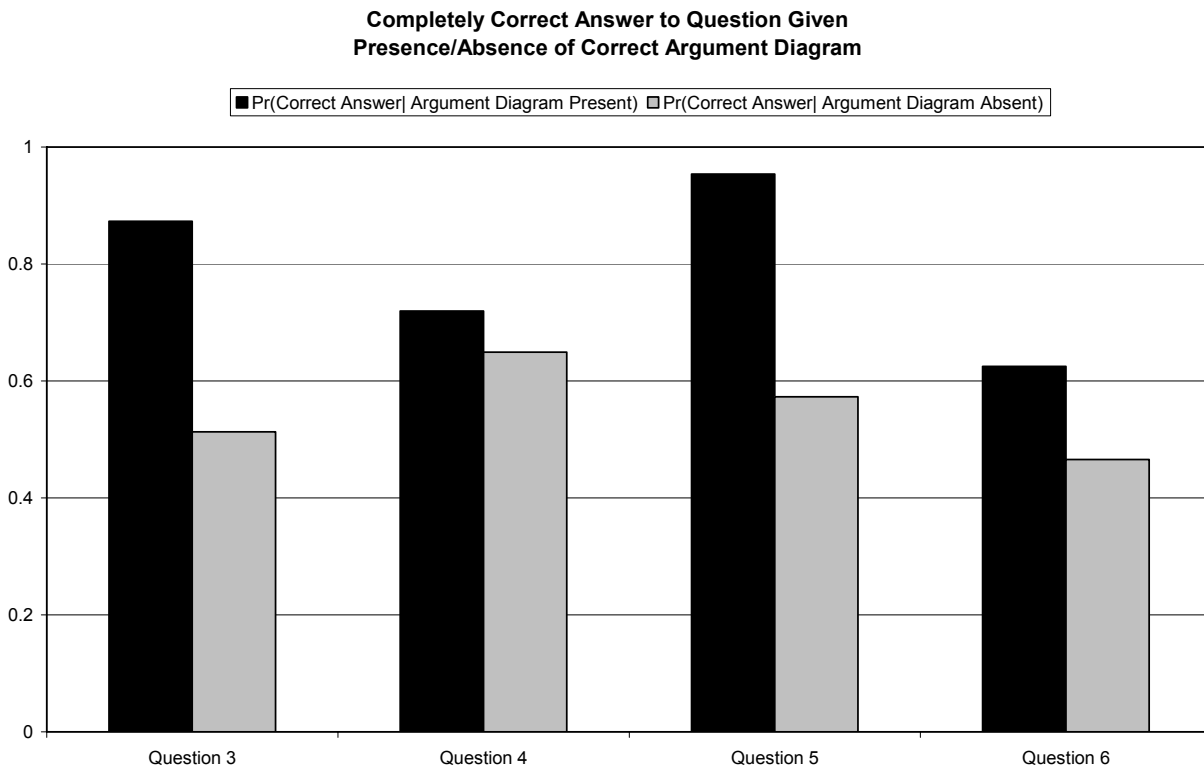


FIGURE 2 Histograms comparing the frequency of students who answered all parts of each question correctly given that they constructed the correct argument diagram for that question to the frequency of students who answered all parts of each question correctly given that they did not construct the correct 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.

*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 the students in Section 1 were the only ones explicitly taught the use of argument diagrams, and all of the students were able to choose their section, 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 section (Section 1, Section 2, Section 3, Section 4) that each had value 1 if the student was enrolled in that section, and 0 if the student was not. 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, Section 1, Section 2, Section 3, Sex, Honors, Grade, and Year as regressors. (Section 4 was omitted for a baseline). The results of these regressions are shown in Table 7.

TABLE 7  
Prediction of posttest, gain, and standardized gain: coefficient (SE coefficient)

	<b>Post</b>	<b>Gain</b>	<b>GainSt</b>
<b>Constant</b>	0.671 (0.052)***	0.671 (0.052)***	1.171 (0.149)***
<b>Pre</b>	0.265 (0.066)***	0.735 (0.066)***	-1.044 (0.189)***
<b>Section 1</b>	0.065 (0.025)**	0.065 (0.025)**	0.133 (0.071)
<b>Section 2</b>	-0.075 (0.026)**	-0.075 (0.026)**	-0.210 (0.074)**
<b>Section 3</b>	0.015 (0.025)	0.015 (0.025)	0.024 (0.070)
<b>Sex</b>	-0.016 (0.019)	-0.016 (0.019)	-0.039 (0.054)
<b>Honors</b>	0.016 (0.027)	0.016 (0.027)	0.025 (0.076)
<b>Grade</b>	-0.022 (0.013)	-0.022 (0.013)	-0.057 (0.038)
<b>Year</b>	-0.004 (0.009)	-0.004 (0.009)	-0.004 (0.025)

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

Next, we performed three more linear regressions—again on the posttest fractional score, the gain, and the standardized gain—this time using PostAD as a regressor, in addition to the pretest fractional score, Section 1, Section 2, Section 3, Sex, Honors, Grade, Year. (Section 4 was again omitted for a baseline). The results are shown in Table 8.

These results show that in both cases a student's pretest score was a highly significant predictor of the posttest score, gain, and standardized gain. 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.

TABLE 8  
Prediction of posttest, gain, and standardized gain: coefficient (SE coefficient)

	Post	Gain	GainSt
<b>Constant</b>	0.672 (0.051)***	0.672 (0.051)***	1.174 (0.146)***
<b>Pre</b>	0.223 (0.066)***	-0.777 (0.066)***	-1.046 (0.189)***
<b>Section 1</b>	-0.013 (0.035)	-0.013 (0.035)	-0.058 (0.102)
<b>Section 2</b>	-0.082 (0.025)**	-0.082 (0.025)**	-0.228 (0.073)**
<b>Section 3</b>	-0.030 (0.028)	-0.030 (0.028)	-0.086 (0.081)
<b>Sex</b>	-0.007 (0.019)	-0.007 (0.019)	-0.015 (0.054)
<b>Honors</b>	0.0004 (0.0264)	0.0004 (0.0264)	-0.016 (0.076)
<b>Grade</b>	-0.020 (0.013)	-0.020 (0.013)	-0.052 (0.037)
<b>Year</b>	-0.0003 (0.0106)	-0.0003 (0.0106)	0.004 (0.025)
<b>PostAD</b>	0.032 (0.011)**	0.032 (0.011)**	0.787 (0.031)*

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

We also see that in both cases 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 sections, as we can see from Table 4. Thus, there seems to be a correlation between which section a student enrolled in and his or her score on the pretest. 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 section the student enrolled in 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.

These results also show that the variables Sex, Honors, Grade, and Year were not significant in either case in predicting a student's posttest score, gain, or standardized gain. In addition, in both cases, the variable Section 3 was not significant as a predictor, which means that the students in section 3 were not significantly different from the students in Section 4, which was taken as the baseline.

In sum, ignoring the variables that were not significant in either table, the two regression equations for each predicted variable can be represented as follows:

*Posttest.*

$$y = 0.671 + 0.265 \text{ Pre} + 0.065 \text{ Section1} - 0.075 \text{ Section2}$$

(0.052) (0.066) (0.025) (0.026)

$p < .001$   $p < .001$   $p = .010$   $p = .005$

$$y = 0.672 + 0.223 \text{ Pre} - 0.013 \text{ Section1} - 0.082 \text{ Section2} + 0.032 \text{ PostAD}$$

(0.051) (0.066) (0.035) (0.025) (0.011)

$p < .001$   $p = .001$   $p = .722$   $p = .002$   $p = .003$

*Gain.*

$$y = 0.671 + 0.735 \text{ Pre} + 0.065 \text{ Section1} - 0.075 \text{ Section2}$$

(0.052) (0.066) (0.025) (0.026)

$p < .001$   $p < .001$   $p = .010$   $p = .005$

$$y = 0.672 - 0.777 \text{ Pre} - 0.013 \text{ Section1} - 0.082 \text{ Section2} + 0.032 \text{ PostAD}$$

(0.051) (0.066) (0.035) (0.025) (0.011)

$p < .001$   $p < .001$   $p = .722$   $p = .002$   $p = .003$

*Standardized Gain.*

$$y = 1.171 - 1.044 \text{ Pre} + 0.133 \text{ Section1} - 0.210 \text{ Section2}$$

(0.052) (0.066) (0.025) (0.026)

$p < .001$   $p < .001$   $p = .062$   $p = .005$

$$y = 1.174 - 1.146 \text{ Pre} - 0.058 \text{ Section1} - 0.228 \text{ Section2} + 0.079 \text{ PostAD}$$

(0.051) (0.066) (0.102) (0.073) (0.031)

$p < .001$   $p = .001$   $p = .573$   $p = .002$   $p = .012$

Here we can clearly see that, before we introduced the variable PostAD, the coefficient for Section 1 was significantly positive for predicting a student's posttest score and gain, and nearly significant ( $p = 0.062$ ) for predicting a student's standardized gain, while the coefficient for Section 2 was significantly negative for predicting a student's posttest score, gain and standardized gain.

After we introduce the variable PostAD, however, the variable Section 1 is no longer significant as a predictor; that is, when controlling for how many correct argument diagrams a student constructed, the students in section 1 were not significantly different from the students in sections 3 and 4. Interestingly, though, the coefficient for Section 2 was still significantly negative for predicting a student's posttest score, gain, and standardized gain, implying that even controlling for how many correct argument diagrams a student constructed, the students in section 2 did worse than students in the other sections. We do not currently have an explanation for this result.

In addition to Section 1 no longer being a predictor, the coefficient for PostAD is significantly positive for predicting a student's posttest score, gain, and standardized gain. This

implies that in fact the only measured factor that contributed to a student's gain from pretest to posttest was his or her ability to construct correct argument diagrams on the posttest.

Thus the instructor for Section 1 was not a contributing factor to the posttest score, gain or standardized gain. Rather, the only factor that does contribute, aside from pretest score, is whether the student constructed correct argument diagrams on the posttest. In other words, regardless of the instructor and the student's personal history, the more correct argument diagrams were constructed on the posttest, the more was gained from the pretest to the posttest.

A stronger version of our second hypothesis, then, is confirmed: constructing correct argument diagrams not only positively contributes to the improvement of argument analysis, but also overrides 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 be 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.

#### **Acknowledgements**

I would like to thank Ryan Muldoon and Jim Soto for their work on coding the pretests 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, Marsha Lovett, and Richard Scheines for detailed comments on many drafts.

### References

- Annis, D., & Annis, L. (1979) Does philosophy improve critical thinking? *Teaching Philosophy*, 3, 145-152.
- 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. (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 (Eds.), *Meeting at the crossroads: proceedings of the 18<sup>th</sup> 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, unlike 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 20<sup>th</sup> 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?