The Improvement of Critical Thinking Skills in *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 to test whether they were actually learning these skills. In addition, there were four sections of this course in the Spring of 2004, and the students in Section 1 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. So, we were also interested in whether this tool would help the students develop the skills we hoped they would master in this course. In each section of this course, 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 in Section 1 gained significantly more than the other students, 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.

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 how sophisticated students' critical thinking skills are or on the most effective methods for improving students' critical thinking skills. The research that has been done shows that the population in general has very poor skills (Perkins, et al., 1983; Kuhn, 1991; Means & Voss, 1996), and that very few courses that advertise that they improve students' skills actually do (Annis & Annis 1979; Pascarella, 1989; Stenning 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. But we want to determine not only whether they are improving, but how much improvement can be achieved using alternative teaching methods.

One of the 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

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

arguments, both one's own and those of others. If we think of an argument they 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, at the end of *Meno*, Plato 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), it is immediately clear how the argument is supposed to work and where we may critique or applaud it.

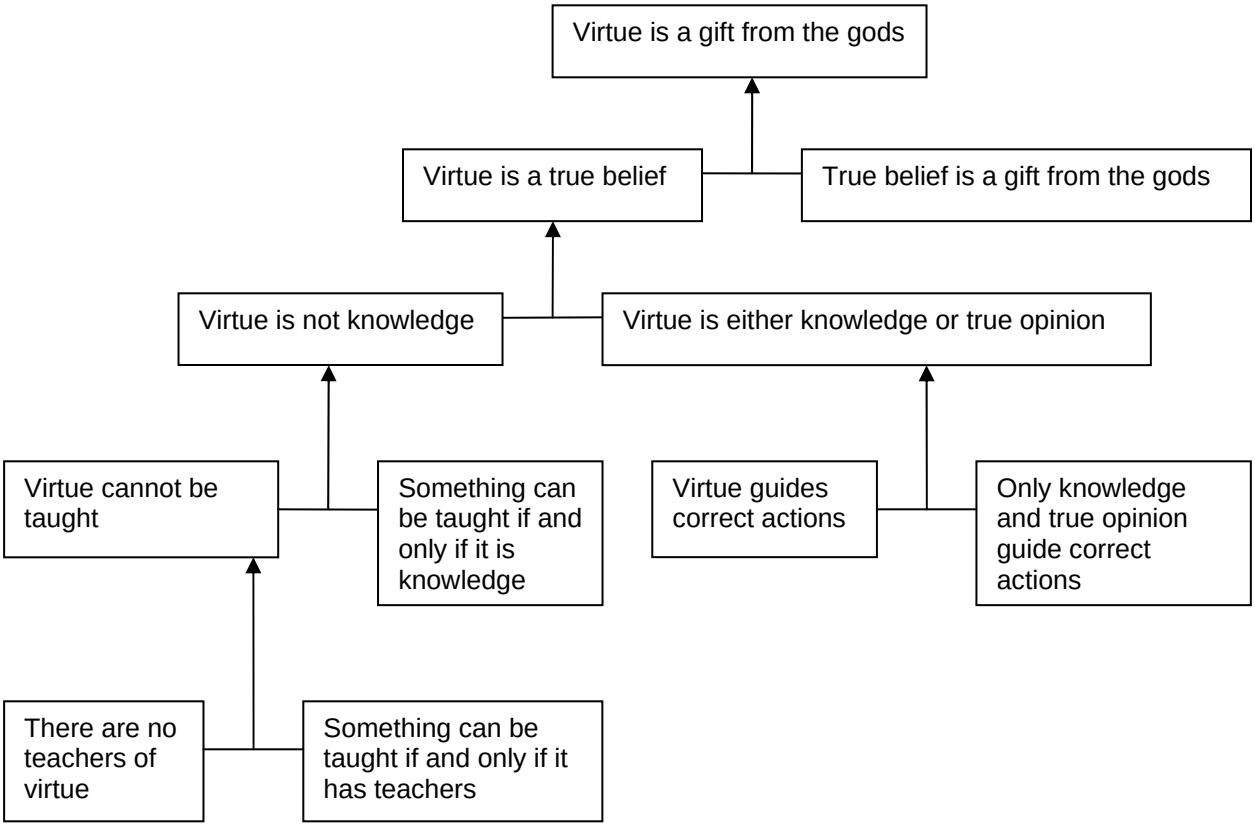

FIGURE 1 An argument diagram representing one of the arguments in Plato's *Meno*.

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; 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 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 to not have this ability. Carnegie Mellon University's introduction to philosophy course (*80-100 What Philosophy Is*), was a natural place to do a semi-controlled experiment. We typically teach 4 or 5 sections of this course 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, it has been determined that it is a topics based course in which epistemology, metaphysics, and ethics will be 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 come from all over the University; they are a mix of all classes and all majors. This study tests this second hypothesis by comparing the pre-test and post-test scores of students in 80-100 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 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.

B. 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 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 "pre-test" (given in Appendix A) and created a companion "post-test" (given in Appendix B). 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.

C. Procedure

Each of the four sections of 80-100 was a Monday/Wednesday/Friday class. The pre-test was given to all students during the second day of class (i.e., Wednesday of the first week). The students in sections 1 and 4 were given the post-test as one part of their final exam (during exam week). The students in sections 2 and 3 were given the post-test on the last day of classes (i.e., the Friday before exam week).

3. Results and Discussion

A. Test Coding

Pre- and post-tests were paired by student, and single-test students were excluded from the sample. There were 139 pairs of tests. 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-/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. 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 1).

TABLE 1
Inter-coder Reliability: Percentage Agreement (PA),
Cohen's Kappa (κ), and Krippendorff's Alpha (α) for each test

	PA	κ	α
Pre-test	0.85	0.68	0.68
Post-test	0.85	0.55	0.54

As this table shows, the inter-coder reliability was fairly good. Upon closer examination, however, 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 pre-test, 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 post-test, 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.

The primary variables of interest were the total pre-test and post-test scores for the 18 question-parts (expressed as a percentage correct of the 18 equally weighted question-parts), and the individual averages scores for each question on the pre-test and the post-test. 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.

TABLE 2
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
A*	Averaged score (or code) on the pre-test for question *
B*	Averaged score (or code) on the post-test for question *
Section	Enrolled section
Sex	Student's sex
Honors	Enrollment in Honors course
Grade	Final Grade in the course
Year	Year in school

B. Gain from Pre-test to Post-test

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 pre-test to post-test was significantly positive. The straight gain, however, may not be fully informative if many students had fractional scores of close to 1 on the pre-test. Thus, our hypothesis can also be 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 pre-test and the post-test, 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 pre-test and the post-test, mean gain (standard deviation), and mean standardized gain (standard deviation)

	Pre	Post	Gain	GainSt.
Whole Population	0.5925 (0.1440)	0.7830 (0.1213)	0.1904 (0.0117)	0.4314 (0.0267)

The difference in the means of the pre-test and post-test scores was significant (paired t-test; $p < 0.001$). In addition, the mean gain was significantly different from zero (1-sample t-test; $p < 0.001$) and the mean standardized gain was significantly different from zero (1-sample t-test; $p < 0.001$). From these results we can see that our first hypothesis is confirmed: overall the students did have significant gains and standardized gains from pre-test to post-test.

An alternate hypothesis for the significant gains, however, is that students improved simply because repeated exposure to a similar test. This hypothesis would seem to imply that the score that a student received on an individual question on the pre-test would be correlated with the score that student received on the corresponding question on the post-test. We tested the Pearson correlation between the code on a question on the pre-test and the code on the corresponding question on the post-test. The results are given in Table 4.

TABLE 4
Correlation between the score on each question on the pre-test with the score on the corresponding question on the post-test

Question	Correlation	*p*-value
1	0.292	0.000
2	0.255	0.002
3a	0.081	0.341
3b	0.029	0.737
3c	0.167	0.050
3e	0.134	0.115
4a	-0.028	0.746
4b	0.091	0.289
4c	0.117	0.172
4e	0.209	0.014
5a	0.073	0.393
5b	0.116	0.173
5c	-0.053	0.536
5e	0.262	0.002
6a	0.095	0.264
6b	0.200	0.018
6c	-0.039	0.646
6e	0.250	0.003

We see here that the score a student received on questions 1, 2, 4e, 5e, 6b and 6e on the pre-test are somewhat correlated with the score that student received on the corresponding questions on the post-test. There are many reasons that could explain this correlation other than mere familiarity with the test. First, nearly every student gave the correct answers to questions 1 and 2 on both tests; thus, they are bound to be correlated. Second, question (e) on each test was the most difficult part of the question—it asked students to evaluate the argument given, and we only coded the answer as correct if the student gave only statements about truth of the premises and quality of support of the premises for the conclusion as reasons for his or her evaluation. This is a very high-level task, and it is reasonable to assume that the students who master this skill do not lose it. Thus if there were a significant number of students who had this skill at the beginning of the class and retained it throughout the semester, and if there were also a significant number of students who did not have this skill at the beginning of the semester or at the end, then we would expect the scores on these questions to be correlated even though there might be a significant number of students who did gain this skill over the course of the semester.

Our second hypothesis was that the students who were able to construct correct argument diagrams would gain the most from pre-test to post-test. Since the use of argument diagrams was only explicitly taught in Section 1, we used a student's section as a proxy for being able to

construct argument diagrams. Testing the hypothesis thus corresponds to 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 pre-test was not the same for each section, and if many students had fractional scores close to 1, on the pre-test. Thus, we also test this hypothesis using the standardized gain. The mean scores on the pre-test and the post-test, as well as the mean gain and standardized gain for the sub-populations of students in each section is given in Table 5.

TABLE 5
Mean fractional score (standard deviation) for the pre-test and the post-test, mean gain (standard deviation), and mean standardized gain (standard deviation)

	Pre	Post	Gain	GainSt.
Section 1	0.6429 (0.1354)	0.8540 (0.1025)	0.2111 (0.0240)	0.5140 (0.0671)
Section 2	0.5255 (0.1648)	0.6952 (0.1408)	0.1697 (0.0270)	0.3200 (0.0536)
Section 3	0.5790 (0.1381)	0.7917 (0.0829)	0.2127 (0.0207)	0.4781 (0.0363)
Section 4	0.6254 (0.1047)	0.7968 (0.0901)	0.1714 (0.0199)	0.4238 (0.0450)

An ANOVA on the post-test indicates that the differences between the sections are significant ($df = 3$, $F = 13.58$, $p < 0.001$), but this is to be expected since an ANOVA indicates that the differences between sections on the pre-test are also significant ($df = 3$, $F = 5.24$, $p < 0.002$). What we are really concerned with, though, is how much each student gained, and an ANOVA on the standardized gain indicates that the differences are statistically significant ($df = 3$, $F = 2.68$, $p = 0.049$), while an ANOVA on the gain indicates that the differences between sections are not statistically significant ($df = 3$, $F = 1.05$, $p = 0.375$).

These results suggest that looking at the average gains and standardized gains for each section may be too crude a method for determining whether the students in Section 1 acquired the desired skills better than those in the other sections. Thus, to test the second hypothesis, we need to use a method that will reveal finer-grained details.

C. Prediction of Post-Test Score and Standardized Gain

To test the hypothesis that knowing how to construct argument diagrams improved students' critical thinking skills, we again used a students' section as a proxy for the former ability. To test whether students in Section 1 improved more than those in the other sections, 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 post-test fractional score, a second for the gain, and a third for the standardized gain—using the pre-test 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 6.

TABLE 6
Prediction of post-test, gain, and standardized gain:
coefficient (SE coefficient) * = p<0.05, ** = p<0.01, *** = p<0.001

	Post	Gain	GainSt
Constant	0.67072 (0.05223)***	0.67072 (0.05223)***	1.1712 (0.1490)***
Pre	0.26452 (0.06627)***	-0.73548 (0.06627)***	-1.0439 (0.1891)***
Section 1	0.06457 (0.02485)**	0.06457 (0.02485)**	0.13329 (0.07090)
Section 2	-0.07470 (0.02595)**	-0.07470 (0.02595)**	-0.21013 (0.07404)**
Section 3	0.01492 (0.02459)	0.01492 (0.02459)	0.02395 (0.07015)
Sex	-0.01624 (0.01893)	-0.01624 (0.01893)	-0.03852 (0.05402)
Honors	0.01608 (0.02661)	0.01608 (0.02661)	0.02535 (0.07592)
Grade	-0.02222 (0.01338)	-0.02222 (0.01338)	-0.05659 (0.03817)
Year	-0.003647 (0.008750)	-0.003647 (0.008750)	-0.00415 (0.02497)

A student's pre-test score was a highly significant predictor of the post-test score, gain, and standardized gain. The coefficient of the pre-test was positive when predicting the post-test, as expected; if all the students' scores generally improve from the pre-test to the post-test, we expect the students who scored higher on the pre-test to score higher on the post-test.

What we really want to know, however, is not only that each student improved from the pre-test to the post-test, but *how much* each student improved. From Table 6, we see that the coefficient of the pre-test was negative when predicting gain and standardized gain. In fact, since the score on the pre-test is a part of the value of the gain and standardized gain, it is interesting that the coefficient for pre-test 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 pre-test shows that none of the variables are significant. We believe that this all can be explained by the fact that scores on the pre-test were not evenly distributed throughout the sections, we can see from Table 2. Thus, there seems to be a correlation between which section a student enrolled in and his or her score on the pre-test. So, a plausible explanation for the negative coefficient when predicting gain is that the students who scored the lowest on the pre-test 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 post-test counted as a part of his or her grade in the course, the students who scored the lowest on the pre-test 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 pre-test. Thus, since we are also concluding that there is a correlation between the section the student enrolled in and the score on the post-test, gain, and standardized gain (see below), there are many contributing factors to a student's gain, the score on the pre-test 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 predicting a student's post-test score, gain, or standardized gain. In addition, 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. Interestingly, though, the coefficient for Section 1 was significantly positive for predicting a student's post-test score and gain, and nearly significant ($p = 0.060$) for predicting a student's standardized gain, while the coefficient for Section 2 was significantly negative for predicting a student's post-test score, gain and standardized gain. One explanation for the positive

coefficients for Section 1 is that knowing how to construct an argument diagram actually helps, and the students in Section 1 were the only ones taught this skill. We do not at this time have an explanation for the negative coefficients for section 2.

We can conclude that the students who had the lowest pre-test scores gained the most from pre-test to post test, and also that, controlling for pre-test score, the students in section 1 gained the most, while students in sections 3 and 4 gained approximately the same amount, and students in section 2 gained the least.

While these results seem to confirm our second hypothesis—that students who learned to construct argument diagrams gained the most from pre-test to post-test—it still is the case that the coefficient for section 1 was only nearly significant for predicting standardized gain. A look at the data may provide a clue as to why. Although the students in section 1 were the only students to be explicitly taught how to construct argument diagrams, a number of students from other sections constructed correct argument diagrams on their post-tests. In addition, a number of the students in Section 1 constructed *incorrect* argument diagrams on their post-tests. 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 is was the other teaching methods of Instructor 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 post-test. 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 post-test was correlated to the student's pre-test score, post-test score, gain and standardized gain. Since there were very few students who constructed exactly 2 correct argument diagrams on the pos-test, 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 post-test. The results are given in Table 7.

TABLE 7
Mean fractional score (standard deviation) for the pre-test and the post-test, mean gain (standard deviation), and mean standardized gain (standard deviation)

	Pre	Post	Gain	GainSt.
No Correct	0.5580 (0.1564)	0.7426 (0.1170)	0.1846 (0.0200)	0.3923 (0.0343)
Few Correct	0.5720 (0.1303)	0.7469 (0.1215)	0.1749 (0.0193)	0.3717 (0.0447)
Many Correct	0.6590 (0.1274)	0.8771 (0.0616)	0.2181 (0.0213)	0.5558 (0.0552)

An ANOVA on the post-test indicates that the differences between the students who constructed 0, Few or Many correct argument diagrams on the post-test are significant ($df = 4$, $F = 22.09$, $p < 0.001$). But, again, what we are really concerned with is how much each student gained. An ANOVA on the standardized gain indicates that the differences are significant ($df = 4$, $F = 4.68$,

$p < 0.012$), while an ANOVA on the gain indicates that the differences are not statistically significant ($df = 4$, $F = 1.19$, $p < 0.307$).

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 post-test 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.

These results, along with the above explanation, also suggest that the ability to construct argument diagrams may not be the only factor contributing to high post-test scores and large gains. To determine what other factors may be present, we performed three more linear regressions—again on the post-test fractional score, the gain, and the standardized gain—using the pre-test fractional score, Section 1, Section 2, Section 3, Sex, Honors, Grade, Year, and PostAD as regressors. (Section 4 was omitted for a baseline). The results are shown in Table 8.

TABLE 8
Prediction of post-test and standardized gain:
coefficient (SE coefficient) * = $p<0.05$, ** = $p<0.01$, *** = $p<0.001$

	Post	Gain	GainSt
Constant	0.67187 (0.05071)***	0.67187 (0.05071)***	1.1740 (0.1459)***
Pre	0.22320 (0.06583)***	-0.77680 (0.06583)***	-1.0460 (0.1894)***
Section 1	-0.01263 (0.03540)	-0.01263 (0.03540)	-0.0576 (0.1019)
Section 2	-0.08200 (0.02532)**	-0.08200 (0.02532)**	-0.22818 (0.07284)**
Section 3	-0.02960 (0.02816)	-0.02960 (0.02816)	-0.08615 (0.08104)
Sex	-0.00657 (0.01867)	-0.00657 (0.01867)	-0.01461 (0.05372)
Honors	0.00044 (0.02643)	0.00044 (0.02643)	-0.01551 (0.07604)
Grade	-0.02027 (0.01301)	-0.02027 (0.01301)	-0.05178 (0.03743)
Year	-0.000321 (0.01068)	-0.000321 (0.01068)	0.00407 (0.02466)
PostAD	0.03183 (0.01068)**	0.03183 (0.01068)**	0.7870 (0.03073)*

These results show that a student's score on the pre-test was still a significant predictor of that student's post-test score, gain, and standardized gain. Here again, the coefficient for the pre-test was positive when predicting the score on the post-test. This is to be expected; regardless of ability to construct argument diagrams, the students who score higher on the pre-test are likely to score higher on the post-test. And again, the coefficient for the pre-test was negative when predicting gain and standardized gain. This is also to be expected; students who scored the lowest on the pre-test gained the most at least because there is more room for them to improve and, since the grade a student received on the post-test counted as a part of his or her grade in the course, the students who scored the lowest on the pre-test 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 pre-test.

Again, the variables Sex, Honors, Grade, Year, and Section 3 were not significant in predicting a student's post-test score, gain, or standardized gain. In addition, 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 Sections 1 and 3 were not significantly different from the students in Section 4, which was taken as the baseline. Interestingly, though, the coefficient for Section 2 was still significantly negative for predicting a student's post-test 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.

The most dramatic result, however, is that the coefficient PostAD is significantly positive for predicting a student's post-test score, gain, and standardized gain. This seems to imply that no matter what the reasons were for a student earning a certain score on the pre-test, those factors were cancelled out by his or her ability to construct correct argument diagrams on the post-test (with the exception of students in section 2). The more correct argument diagrams a student constructed on the post-test, the more the student gained from the pre-test to the post-test.

Thus it seems that our second hypothesis is also confirmed: all else being equal, students who learned to construct argument diagrams gained more from pre-test to post-test than those who did not.

D. Prediction of Score on Individual Questions

The hypothesis that students who were able to construct argument diagrams improved their critical thinking skills the most can also be 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 post-test on (b) the student's ability to answer the other parts of that question correctly. The hypothesis implies 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.

A Pearson correlation was run for each question and the latter set of new variables against the corresponding variable in the former set of new variables. The results are given in Table 9.

TABLE 9
Correlation of scores on question-parts and completely correct question scores with constructing the correct argument diagram on that question

Question/Total	Correlation	P-Value
3a	0.083	0.334
4a	-0.041	0.635
5a	0.158	0.063
6a	0.065	0.449
3b	0.313	0.000
4b	0.158	0.064
5b	0.416	0.000
6b	0.343	0.000
3c	0.276	0.001
4c	0.180	0.034
5c	0.022	0.796
6c	-0.061	0.477
3e	0.392	0.000
4e	0.154	0.070
5e	0.395	0.000
6e	0.185	0.029
3Tot	0.383	0.000
4Tot	0.075	0.381
5Tot	0.381	0.000
6Tot	0.074	0.385

These results show that constructing the correct argument diagram was not correlated with the score on part (a) of each question. This is to be expected as part (a) was a very easy question for the students in general; almost all the students answered part (a) of each question correctly. Constructing a correct argument diagram was, however, correlated with giving the correct answer to parts (b) and (e), except for question 4, in which is was only marginally correlated. This can be explained by fact that question 4 was by far the easiest question on the test, and students in all sections scored the highest on this question. Constructing the correct argument diagram was correlated to giving the correct answer on part (c) only for questions 3 and 4, and not for questions 5 and 6. This is seemingly an odd result, as argument diagrams should be particularly helpful in determining the relationship of the premises to each other and the conclusion. However, these results can be explained by the fact that there was some confusion in the wording of the question. The question asked the students to circle the *one* correct answer on part (c) of each question, when the correct answer for part (c) on questions 5 and 6 was actually a combination of answers. Thus, when coding the test, the coders were instructed to assign 1 to any answer that was any part of the correct answer. Thus, it was very easy for a student to be coded as answering part (c) of questions 5 and 6 correctly, and this is reflected in the fact that these two parts were among the highest scoring on the post-test.

These results also show that constructing the correct argument diagram was correlated to giving the correct answers to all parts of questions 3 and 5, but not to giving correct answers to questions 4 and 6. This can be explained by the fact that, as previously mentioned, question 4 was an easy question, and by the fact that question 6 was an exceptionally hard question. Almost no students gave the correct answers to all parts of question 6, and so there is not enough data to tell whether constructing the correct argument diagram aided these students. It is instructive to

see the correlations between constructing the correct argument diagram and answering correctly all parts of each questions; thus, I have included the histograms in Figure 2.

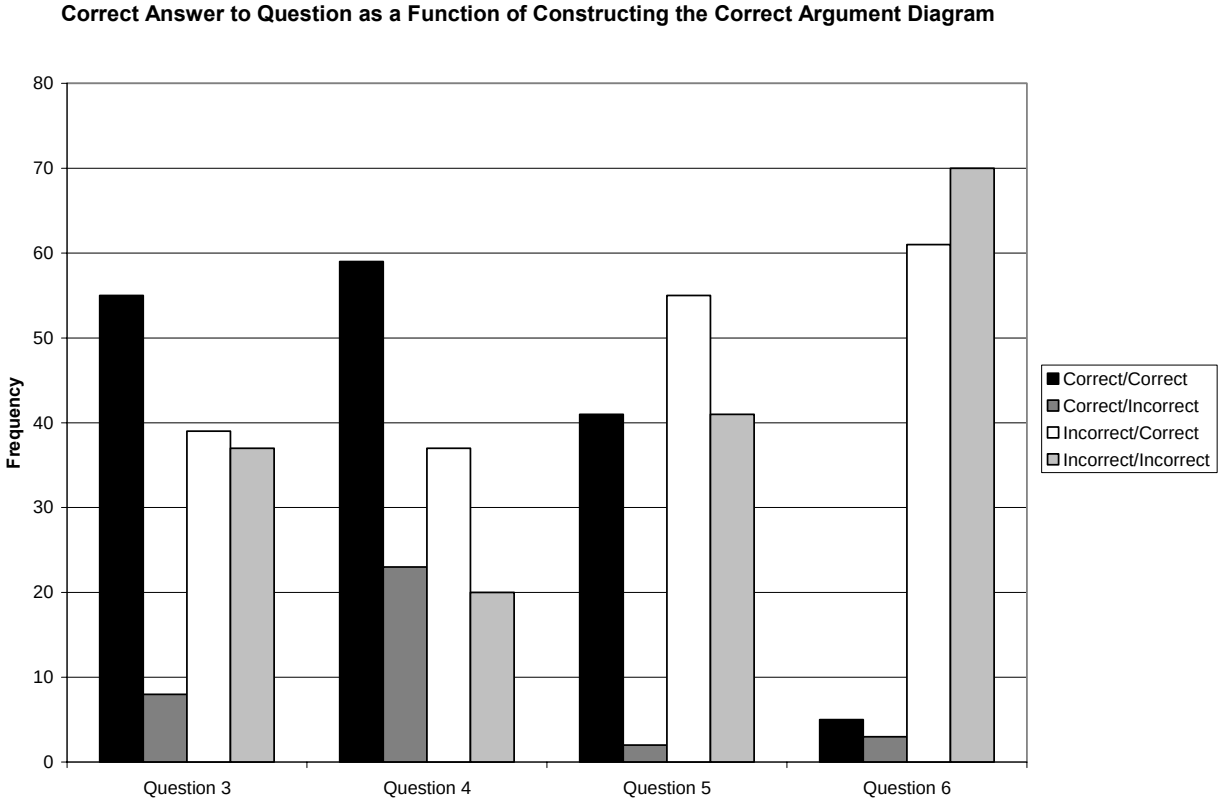

FIGURE 2 Histograms showing frequency of students who did or did not answer all parts of each question correctly, divided by whether they constructed the correct argument diagram for that question. In the legend, the first term indicates whether the student constructed the correct argument diagram; the second term indicates whether the student answered all parts of the question correctly.

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 seem to 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.

5. 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.

The use of argument diagrams is supposed to aid students in all 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, there is much less ambiguity in deciding on what bases to evaluate the argument.

Since the scores on part (a) of each question were high on the pre-test, and even higher on the post-test, we conclude 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. However, it 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 learned how to construct argument diagrams. Constructing the correct argument diagram was highly correlated 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 pre-test 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.

6. 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.

7. Future Work

This study raises many more questions than it answers. While it seems 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 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.

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.

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.

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?