Analysis of CSU Assessment Critical Thinking Data

Jamel Ostwald, History

Methodology

Six faculty members teaching eight separate classes participated in the study, assessing critical thinking in both 100-level and 200-level courses in a variety of disciplines. Some of the faculty were assigned to the control group, while others were assigned to the experimental group that was to use argument mapping. Two of the faculty on the Critical Thinking Workgroup also participated, putting one of their sections in the control group and another in the experimental group. The summary data follows:

Figure 1: Summary of Classes Participating in Study

Class	Level	N pre	N post	N change	C/E[1]	Size cap
A	200	31	19	19	Experiment	35
B	100	20	15	15	Experiment	20
C	200	13	12	12	Control	20
D	200	11	11	11	Experiment	20
E	100	24	20	20	Experiment	25
F	200	19	19	16	Control	25
G	200	16	14	14	Control	25
H	200	9	8	8	Experiment	25
TOTAL		143	118			

Our initial intention was to find volunteers in the First-Year Program so as to measure the effectiveness of argument mapping in teaching first-year students critical thinking. As it turns out, this plan had to be modified in two key respects. First, we were unable to find enough volunteers at the 100-level, so we enlisted faculty who were teaching classes at the 200-level as well. Second, and more importantly, as the Post-Evaluation Survey section below details, faculty who volunteered for the experimental group did not implement argument mapping beyond a basic introduction to the procedure and a few attempts to incorporate it into their courses. Therefore, our results can only address the issue of how existing faculty attempts to teach critical thinking were or weren't successful in lower-level courses.

We assigned two scorers to score each student's pre-test and post-test, making sure that faculty on the Critical Thinking Workgroup did not score their own classes. After designing the pre-test and post-test based off of relatively simple argumentative passages, we spent an early meeting sharing a scoring rubric and calibrating our model answers on the pre-test. The difference between the scorers for both pre-test and post-test averaged approximately 0.5 points (median of 0.5), and in the 8 cases where the two scores disagreed by 2 points or more, a third scorer was brought in. We then averaged the scores to create an average score for both the pre-test and post-test for each student – Pre-Ave

[1] The difference between control group and experimental group turned out to be irrelevant, as faculty did not implement the argument mapping process to any significant extent in their experimental sections.

and Post-Ave. The pre-test averaged score was subtracted from the post-test averaged score to create the Change variable. We also measured the variation between our two scores in the Pre-Reliability and Post-Reliability variables.

Boxplots are visual displays of the distribution – the bottom line indicates the bottom quartile (25%), the lower part of the box the 2nd lowest quartile (i.e. values falling in the 26%-49% range), the horizontal line dividing the box is the median, the top part of the box is the 3rd quartile (values 51%-75%), and the top line indicates the top quartile (76%-100%). Significant outliers are indicated by asterisks (*).

For more information on boxplots, start with http://en.wikipedia.org/wiki/Box_plot.

Reliability Between Scorers

In spite of calibration, the following boxplots illustrate the variation between scorers.

Figure 2: Reliability of Pre-Average Score by Class

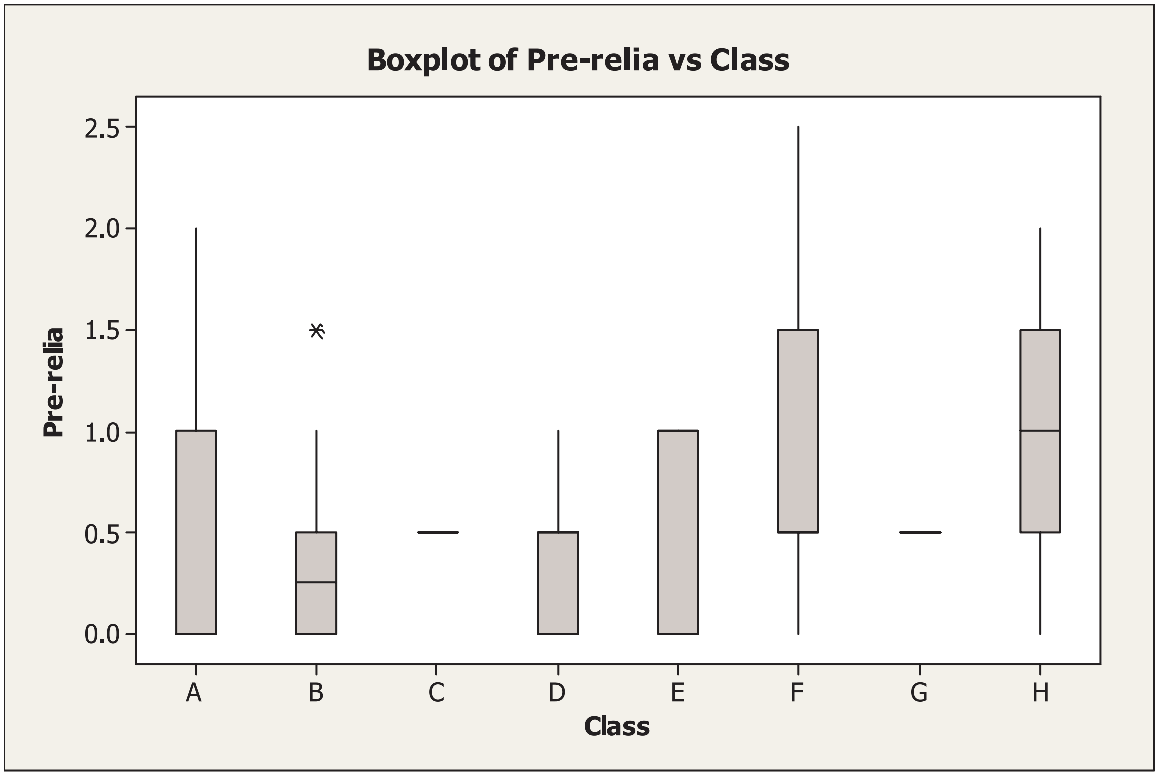
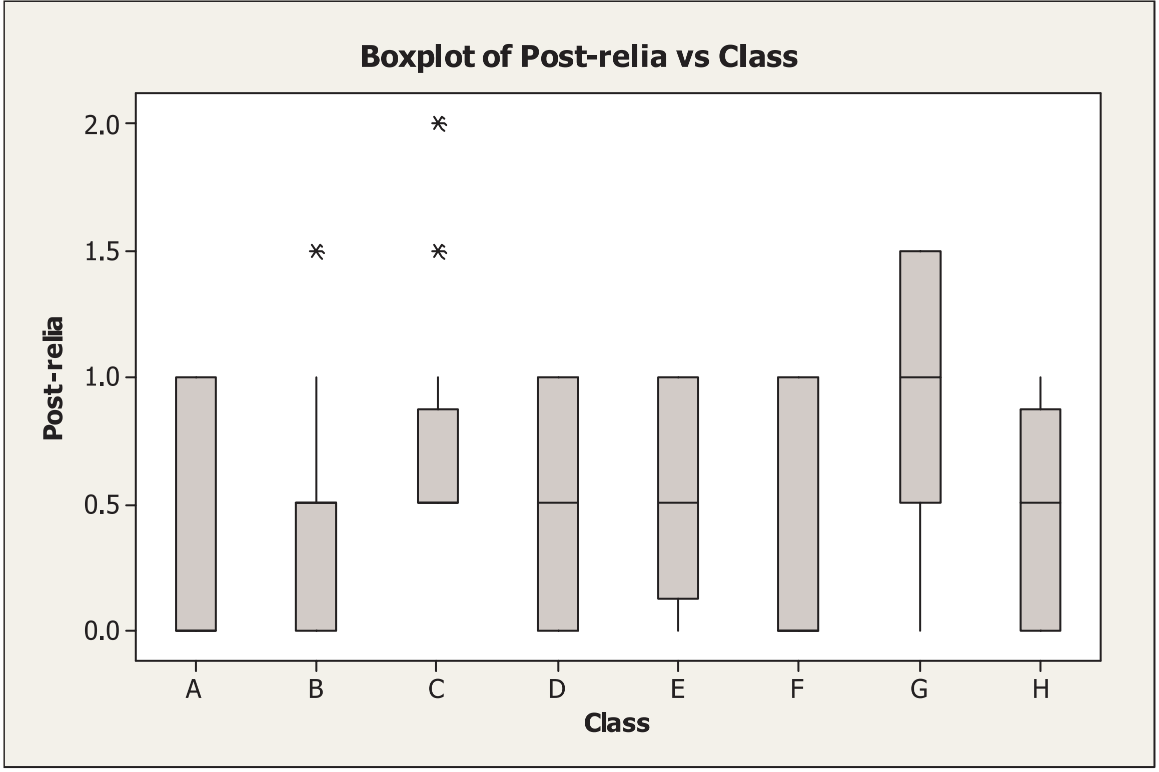


Figure 3: Reliability of Post-Average Score by Class



As both of these charts indicate, there was an average variation of 0.5 between the two scorers assigned to each class. Thus, differences of 0.5 or less are likely not significant, as they are within the 'error.' Future studies should consider keeping the scorer pairings consistent from pre-test to post-test, when the goal is to measure change over the course of the semester.

Pre-Test

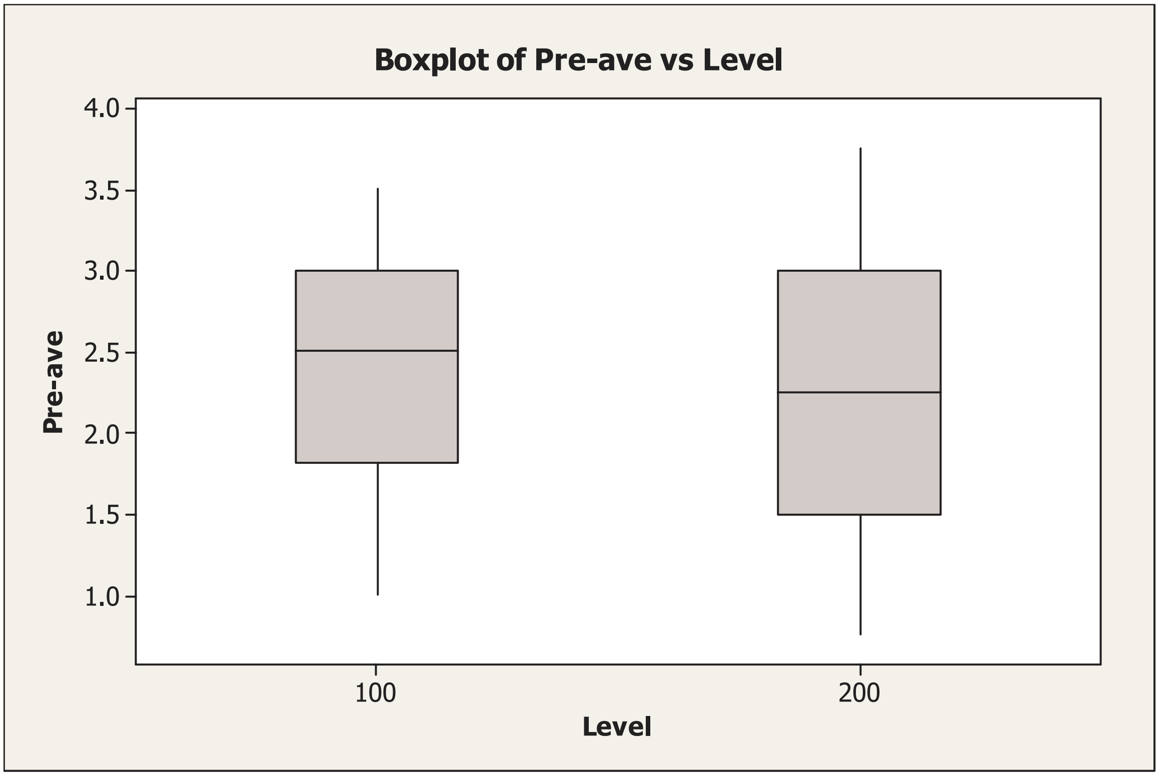
The pre-test data illustrates the wide variation in student abilities to understand even relatively-simple arguments. Below is a stem-and-leaf diagram of the distribution of the averaged pre-test scores, showing a relatively normal distribution around the median of 2.25. The mean averaged pre-test score was 2.3 with a standard deviation of 0.81.

Figure 4: Distribution of Pre-Average Scores

```
N  = 143
Leaf Unit = 0.50
N* = 3


 3    0  777
 23   1  00000000000022222222
 45   1  5555555555555777777777
(29)  2  00000000000000000222222222233
 69   2  5555555555555555677777777778
 41   3  000000000000000000222222
 17   3  55555555555557777
```

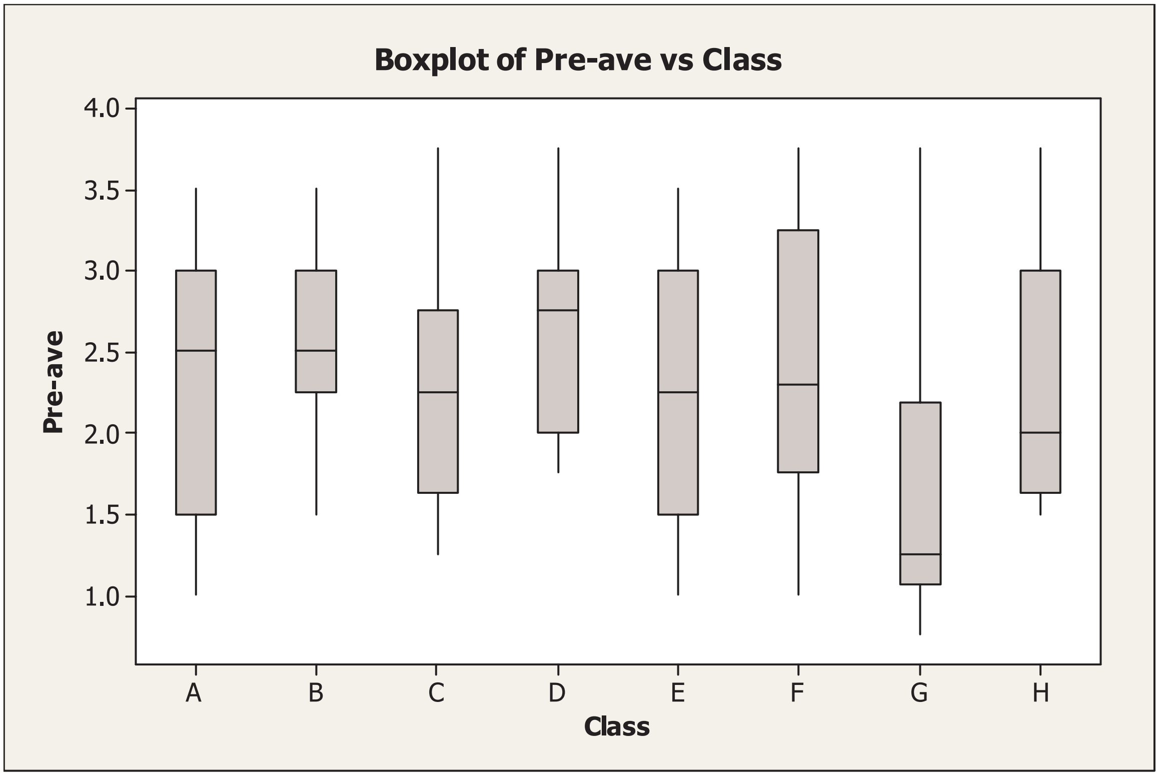
As the following comparative boxplot suggests, students in the 200-level courses tended to score a slight bit worse with the pre-test argument than those in the 100-level courses (a median of 2.25 versus 2.3), although given the variation between scorers, a margin of 0.5 or less is probably not significant.

Figure 5: Comparative Distribution of Pre-Average Score by Level of Course



Here is a comparative boxplot of the distribution of the averaged pre-test scores by course:

Figure 6: Comparative Distribution of Pre-Average Score by Class



As the above indicates, most courses hovered around 2.0 - 2.5 with one significantly lower than the rest (G). Some classes tended to have greater variation (e.g. A, E, F), while most classes had at least a few students scoring well on the pre-test.

<u>Post-Test</u>

A similar post-test was scored in the same manner as the pre-test, with two scorers' scores being averaged into a Post-Ave score.

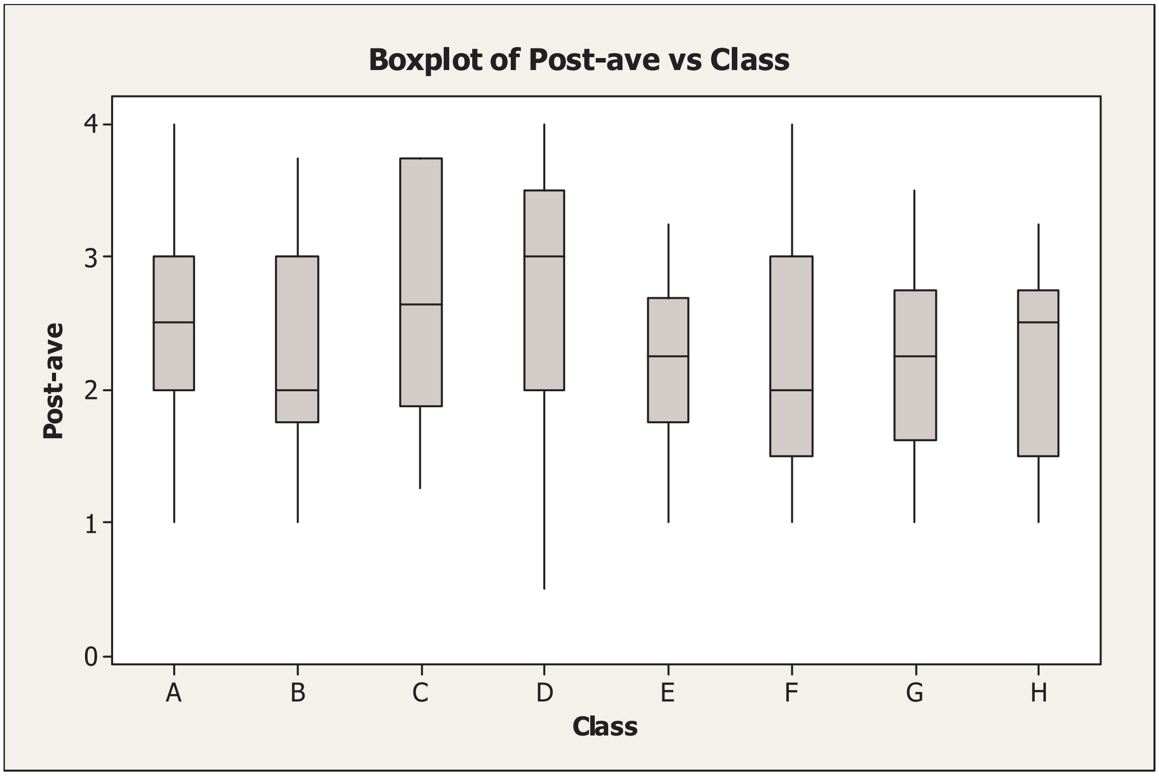
A stem-and-leaf diagram of the averaged post-test scores shows little difference from the pre-test. The mean overall was 2.4 with a standard deviation of 0.85. The reliability between the 2 scorers was comparable to that on the pre-test, although there was a slightly smaller standard deviation.

Figure 7: Distribution of Post-Average Scores

```
N  = 118
Leaf Unit = 0.50
N* = 28


 1    0  5
 14   1  0000000002222
 32   1  555555557777777777
 57   2  0000000000000000002222222
(25)  2  5555555555555555577777778
 36   3  00000000000222
 22   3  55555555577777777
 5    4  00000
```

Here is a boxplot of the distribution of the averaged post-test scores by course:

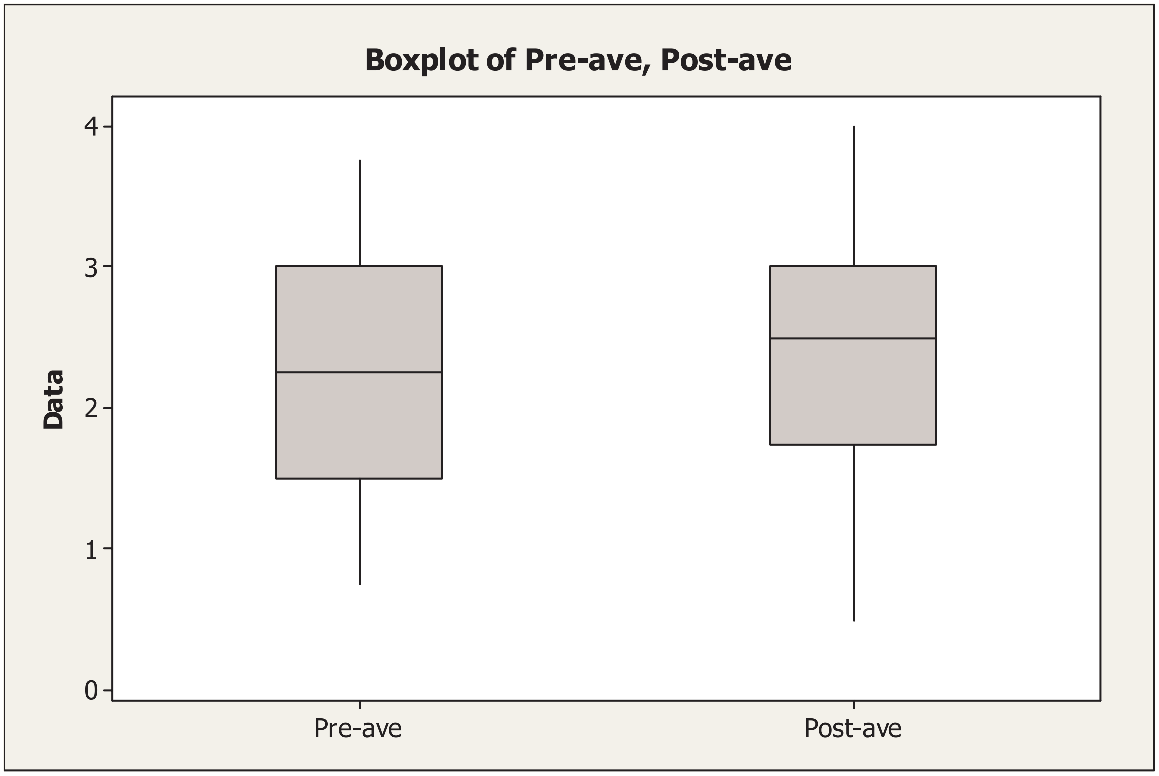
Figure 8: Comparative Distribution of Post-Average Score by Class



In the post-test, the 200-level students did perform slightly better than their 100-level peers, although this too is well within the variation between scorers.

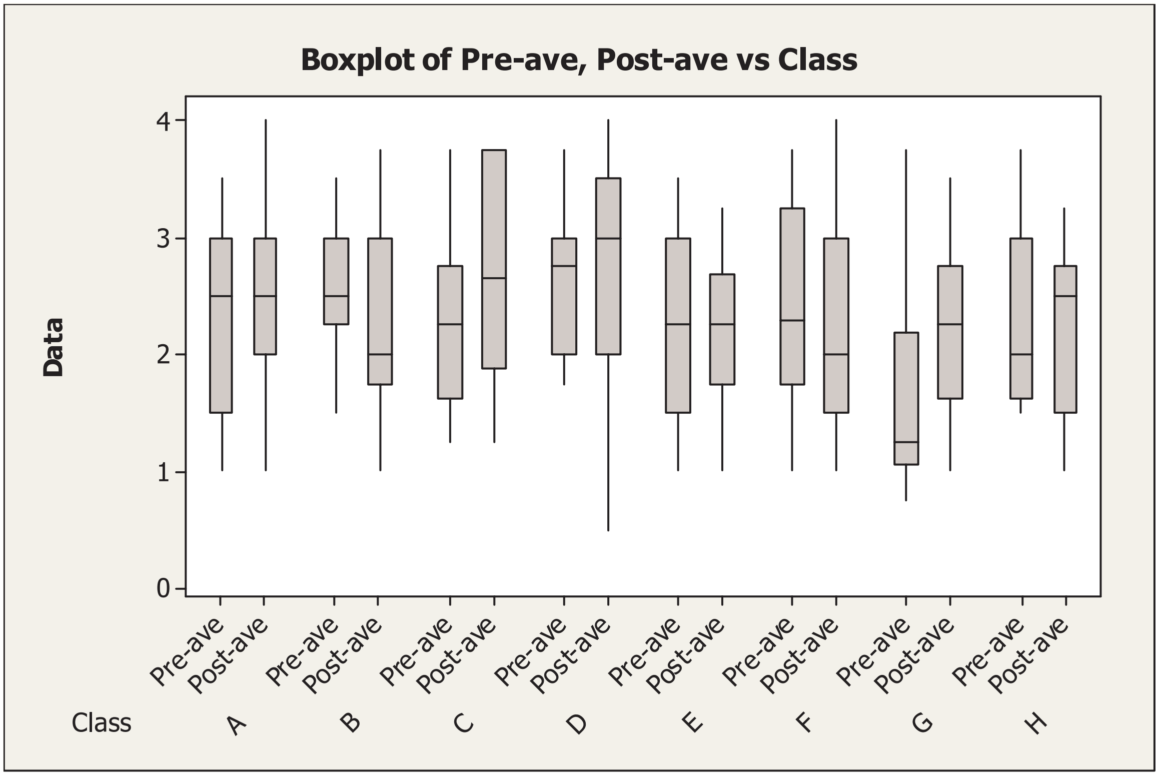
A boxplot comparison of the averaged pre-test scores and post-test scores also suggests very little change overall, in either the central tendencies or the distribution.

Figure 9: Comparative Distribution of Pre-Average & Post-Average



This overall lack of change masks, however, greater variation when we look at the class-level data. Comparing the change in distributions between pre-test and post-test side-by-side, we see:

Figure 10: Comparative Distribution of Pre-Average & Post-Average by Class



As the above chart suggests, there was no general trend (even if we were to distinguish between the controls and experimental groups). A few classes saw no change in the median but a decrease in the variation around the mean (A and E), others saw an increase in the post-test median results with an increase in variation around the median (C and D), one saw an increase in the median with the same variation (H), another saw a decrease in the median post-test score with greater variation than before (B), and one saw a decrease with the same variation (F). In short, there does not seem to be any discernable pattern. Even in the most significant cases, however, the median improvement was not very strong: C improved its median by 0.5 points, whereas class G improved its by 0.7 points. Both of these are within the variation between scorers, and a look at Figure 3 will show that G's post-test had an unusually large difference between the two scorers' results, while C's post-test reliability was also larger than most (especially considering the two outliers). These two classes seem to have done better than the others in improving a significant number of their students (i.e. shifting the majority of their class upward), although the amount of improvement does not appear to have been very substantive when measured in terms of points.

Change in Scores

Another way of comparing students' performance pre-test and post-test is to calculate the change between the pre-test and post-test scores for each student.[2] Overall, the change in

[2] We are limited, of course, in our ability to claim that this specific course was responsible for any improvement in a student's critical thinking.

scores between averaged pre-test and averaged post-test was not dramatic, with a mean change of only 0.15 (median of 0) – this is well within the range between scorers. The standard deviation of 1.1 highlights how some students significantly improved while others worsened, sometimes equally significantly. The following stem-and-leaf diagram of the distribution gives details:

Figure 11: Distribution of Change between Pre-Test and Post-Test

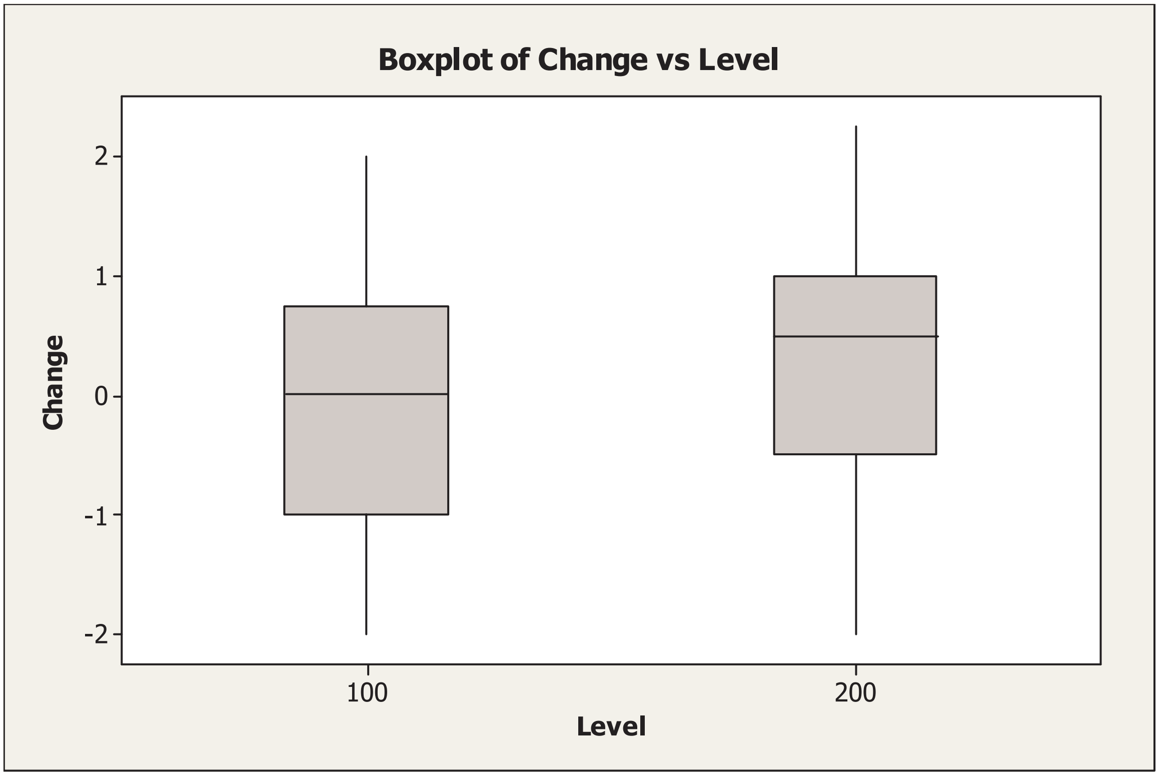
```
N  = 115
Leaf Unit = 0.50
N* = 31


 3   -2  000
 11  -1  77775555
 22  -1  32222200000
 42  -0  77777777755555555555
 51  -0  220000000
(12)  0  000000022222
 52   0  5555555577777777777
 33   1  0000000022222224
 17   1  5555557777
 7    2  0000022
```

Since faculty did not use the argument mapping process extensively in their experimental sections (see Post-Evaluation Survey below), the change data analyzed here is for the full set of students who had both pre-tests and a matching post-test. There were 28 students who took the pre-test but did not take the post-test, or 20% of those taking the pre-test. These 28 pre-test scores were evenly distributed from 1 to 3.75, so these missing students would seem to be representative of the broader student population.

The distribution shows that the median student saw no change from pre-test to post-test, while the number of students improving was equally offset by the number of students scoring worse on the post-test. Seven students improved significantly (+2), whereas several scored much worse (-2).

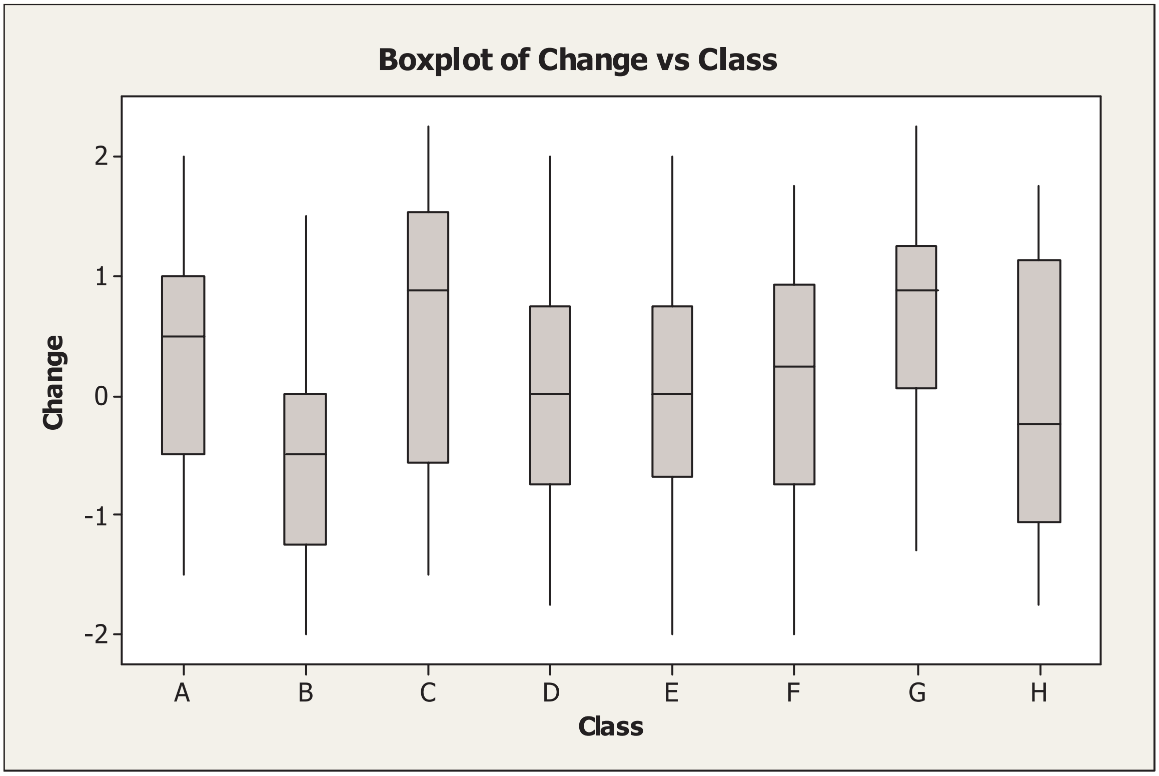
When comparing the change in averaged scores by the level of the course, we see a slight improvement in the 200-level courses versus the 100-level courses (median change of 0.5 versus 0), although here too the 200-level improvement is within the range of the scorers' variation.

Figure 12: Comparative Distribution of Change by Level of Course



Here is a boxplot of the distribution of the average change in scores by course:

Figure 13: Comparative Distribution of Change by Class



Here too we see that the average course saw no significant improvement in its median score, and that as a result their students were as likely to perform worse on the post-test than on the pre-test. The two cases of C and G show a higher improvement than most, and class A's improvement is also noteworthy. It is also noteworthy, however, that in every class at least 25% of the students did worse at the end of the term. Caution is needed with these conclusions, not only because of the variation between scorers, but also since classes with higher pre-test scores were limited in their ability to improve scores for the post-test – we would not, however, expect to see a decrease in scores. Class G, for example, may well have benefited from its students scoring much more poorly on the pre-test than the other sections (Figure 6), allowing for more opportunity for improvement. We should also consider the potentially confounding effect of class size: several of the courses were capped at 20 students (B-D), while A had 35 students. This size and format difference would also need to be taken into account when explaining any divergence from class to class.

<u>Post-Evaluation Survey</u>

A post-evaluation survey was sent to the experimental faculty to gauge their experience and interest in the project. The following table summarizes the result:

Figure 14: Frequency of Survey Response by Question[3]

Survey Question	Not at all	Very little	Some-What	Very much	Extensively
Professor used the process...	2		2		
Professor used the software...	2	2			
Students used the process...		3	1		
Students used the software...	3	1			
Professor found the software useful...	2	1			
Students found the process useful...	3	1			
Likely future use of the process...	2		1		
Process found useful overall...	2		1		
Software found useful overall...	2				

Several points in the table above are worth highlighting:

- The process was rarely used in the classes of the experimental groups ("Very little" in three cases, and "Somewhat" in one). Defining these frequencies was left up to the respondent – each experimental group (except class E) included one of the workgroup members presenting a 30-45 minute introduction to argument mapping at the beginning of the course.
- Two of the four professors in the experimental group reported that they personally used the process "Not at all," while two reported using it "Somewhat."
- The software was used by almost none of the students.
- Two of the three responding faculty saw no use in the argument mapping process.
- One faculty member was responsible for all but one of the "Somewhat" responses.

In short, there was little interest in argument mapping from three of the four experimental faculty.

The only two written comments were that 1) the process was "overly complex" and that "more intuitive methods" could extract the same information; and 2) that the difficulty in teaching critical thinking was with the students' lack of training in argumentation, and that with more time and assistance, this particular professor would be willing to experiment more with the process in future classes.

Conclusions

Keeping in mind the limitations of the study, the above data suggests the following points:

1. Eastern students taking lower-level courses vary widely in their abilities to understand simple prose arguments: on a 4-point scale, averaged scores ranged from 0.75 to 3.75 on the pre-test and this wide variation was still present in the post-test, with scores ranging from 0.5 to 4.0. Median scores were 2.3 on the pre-test and 2.4 on the post-test.

[3] All questions do not add up to 4 due to some respondents omitting answers and/or responding "N/A."

2. Students taking 200-level courses were, overall, no better at understanding arguments than those taking 100-level courses, although they may have been, on average, slightly more likely to see some small improvement over the course of the term.
3. Some classes appear to have done better than others at improving averaged median scores from pre-test to post-test, although the difference does not appear to be very significant given the variation in scoring, and also seems to be related to how well they scored on the pre-test. The impact of smaller class sizes is also worth considering. There was significant variation in the amount of improvement, however, and at least 25% of students in all classes scored *worse* on their post-test than their pre-test. If these tests are an accurate measure of critical thinking skills, the *average* student saw no practical improvement over the course of the term, and was as likely to do worse as do better.
4. Early faculty involvement is crucial – if the faculty do not believe in the process, and/or if the faculty do not have adequate training and time to incorporate it into their syllabi, they will not use it. Further, introducing a process such as argument mapping (or having an outsider do it) and then not implementing it might even do more harm than good.

Limitations of the Study

There were several limitations to the study, which require that the above data be taken with a grain of salt.
1. Most critically, we were unable to test the impact of argument mapping because almost none of the faculty incorporated it into their courses in any substantive way (as revealed in the Faculty surveys). We cannot give much credence to differences between the results of the control and experimental groups, and therefore we have not calculated the differences between these two groups (they were largely insignificant). The most this study can do is show the wide variation between students at the same level of coursework and the general lack of many students' ability to understand the most basic of arguments.
2. The 4-point scale used in both pre-test and post-test is not associated with any particular value or underlying index. As such, it is difficult to assign substantive meaning to specific increases or decreases in scores – we can only judge the relative amount of change.
3. The pre-test and post-test were not calibrated as ideally as might be desired. Although both tests asked students to identify the particular parts of an argument, the specific questions asked were different between the pre-test and the post-test. A few students appear to have misinterpreted what the questions on the post-test were asking, particularly that we wanted the students to provide the *author's* reasons rather than their own, and that we defined reasons *argumentatively* (i.e. evidence for how we know that a particular claim is likely) rather than reasons in the sense of explanations (i.e. an explanation of why a particular claim was true). This was discussed in the introduction to argument mapping, but it is unclear how much it was reinforced during the rest of the term.
4. The mindsets of the students taking these pre-tests and post-tests were unclear. The tests were not part of the students' grades, and in at least a few cases, it was obvious that the post-test was not taken seriously.