CRITICAL THINKING ASSESSMENT REPORT

PREPARED BY BYRON JAVIER, ASST. DEAN OF RESEARCH AND PLANNING
MALCOLM X COLLEGE
APRIL 22, 2009

CRITICAL THINKING ASSESSMENT REPORT

Introduction

As part of the Assessment plan, Malcolm X College faculty engaged during the fall 2008 semester in an extensive process to assess Critical Thinking at the departmental level. A basic principle was to explore the learning process of a value-added education. The process had several components: deciding about a rubric that would measure effectively Critical Thinking, preparing the different instruments that each department would use with their students; collecting and analyzing test results; and, finally, administering the same instrument in class again to determine whether there was a change in students' critical thinking skills.

Faculty at institutions of higher education work explicitly to develop critical thinking among students and indeed, most courses implicitly, if not explicitly, have critical thinking as a goal. Thus, critical thinking is a higher order thinking skill exhibited in context. At the college level, it is learned, developed and finds formal expression within contexts represented by academic disciplines.

A common rubric was selected by the Assessment Committee. The rubric included criteria to address the following components of critical thinking:

- Interpretation/Identification of facts (Criterion 1)
- Argument (Criterion 2)
- Thoughtful Analysis (Criterion 3)
- Evaluate Alternatives (Criterion 4)
- Justification/Explanation of Reasons (Criterion 5)
- Draw Conclusions (Criterion 6)

Description of Methodology

For our purposes, the pre-post test was based using a rubric. Each department designed their own test. Additionally to insure validity of the measurement all faculty members had to agree and teach content focused on the same Student Learning Outcome for the critical thinking aspect of the course. For grading purposes, it was decided that two instructors were to grade each pre-post test using the rubric. In addition, the instructors grading the tests were comprised of one teaching the course and one other faculty person in their department.

The pre-test was administered the first week of school. Instructors provided feedback to students in the form of a score showing how well he or she performed on each aspect detailed in the rubric. Students did not get their test back because the same test would be administered two weeks after their midterm.

The sample population included all of the students enrolled in selected sections. Overall forty-three sections participated in the pre-test data collection: thirty-nine from the General Education division, and thirteen from the Career programs. The departments and programs that participated in this process included: Biology, Business, Chemistry, Child Development, English, Renal, Respiratory Care, Mathematics, Mortuary Science, Nursing, Phlebotomy, Social Science, and Surgical Tech. The total number of students that participated in both the pre- and post-test was 724.

Department/Program	**Number of Sections**
Biology	7
Business	5
Chemistry	8
Child Development	2
English	9
Mathematics	9
Mortuary Science	1
Nephrology/Renal	1
Nursing	1
Phlebotomy	1
Respiratory Care	1
Social Science	6
Surgical Tech	1
Total	**52**

Interrater reliability

Students' responses to the instrument were graded by 2 raters and all written work was rated using a rubric. Since each test was graded by two different raters, we needed to determine interrater reliability. Interrater reliability is concerned with the consistency between the judgments of two or more raters. In the past, interrater reliability has been measured by having two (or more) people who make decisions or ratings across a number of cases and then find the correlation between those two sets of decisions or ratings. That method gives an overestimate of the interrater reliability so we will not use it. For our purposes, we will use the Kappa coefficient. This is one of the more commonly used measures of interrater reliability. Kappa is a measure of agreement. It is currently gaining popularity as a measure of scorer reliability. The results of the interrater analysis are Kappa = 0.676 with $p < 0.001$. This measure of agreement, while statistically significant, is only marginally convincing. As a rule of thumb, values of Kappa from 0.40 to 0.59 are considered moderate, 0.60 to 0.79 substantial, and 0.80 outstanding (Landis & Koch, 1977). Most statisticians prefer for Kappa values to be at least 0.6 and most often higher than 0.7 before claiming a good level of agreement.

Criterion 1: Interpretation/Identification of facts

The proportion of agreements for the pre-test was 40%, with kappa (N = 1,128) = .401, and p < .0005. The proportion of agreements for the post-test was 47%, with kappa (N = 625) = .469, and p < .0005.

Criterion 2: Argument

The proportion of agreements for the pre-test was 29%, with kappa (N = 1,117) = .286, p < .0005. The proportion of agreements for the post-test was 28%, with kappa (N = 626) = .275, p < .0005

Criterion 3: Thoughtful Analysis

The proportion of agreements for the pre-test was 35%, with kappa (N = 1097) = .350, p < .0005. The proportion of agreements for the post-test was 20%, with kappa (N = 619) = .201, p < .0005

Criterion 4: Evaluate Alternatives

The proportion of agreements for the pre-test was 30%, with kappa (N = 1,061) = .300, p < .0005. The proportion of agreements for the post-test was 25%, with kappa (N = 612) = .254, p < .0005

Criterion 5: Justification/Explanation of Reasons

The proportion of agreements for the pre-test was 33%, with kappa (N = 1,033) = .329, p < .0005. The proportion of agreements for the post-test was 21%, with kappa (N = 608) = .308, p < .0005

Criterion 6: Draw Conclusions

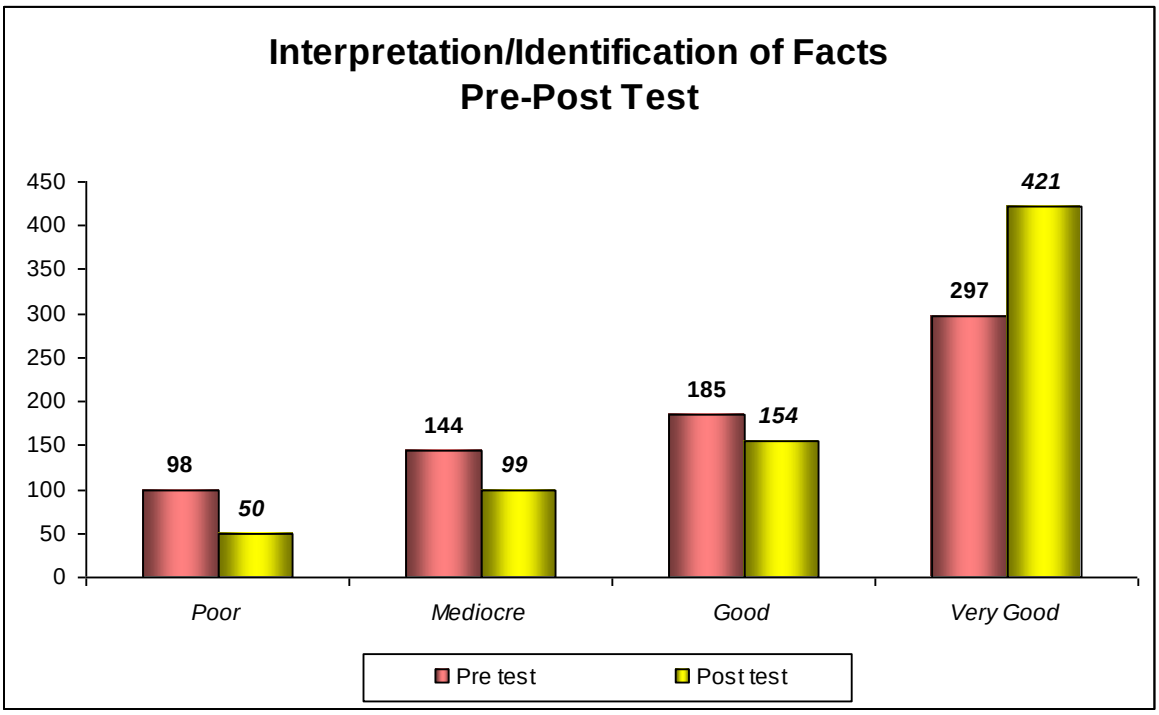
The proportion of agreements for the pre-test was 34%, with kappa (N = 1,033) = .344, p < .0005. The proportion of agreements for the post-test was 30%, with kappa (N = 605) = .301, p < .0005

Analysis and Presentation of Results

The rubric used by all raters included four explicitly defined competence categories for each criterion: Very Good (4), Good (3), Mediocre (2), Poor (1). Whenever a student did not provide a response, raters did not enter a score. The results presented here include the number and percentage of students' work that were scored at each competency level. The data have been aggregated across all courses.

Criterion 1: Interpretation/Identification of facts

	Pre-Test		Post-Test	
	Frequency	Percent	Frequency	Percent
Poor	98	14%	50	7%
Mediocre	144	20%	99	14%
Good	185	26%	154	21%
Very Good	297	41%	421	58%
Total	724	100%	724	100%
Missing	0	0%	0	0%
	724	100%	724	100%



For Criterion 1, the assessment results indicate a shift between the pre and post test for category 1 (Poor). The results show a 49% decrease of students who during the pre-test did poorly to identify correctly the facts of the problem. Results for Category 2 (Mediocre) also show a decline of 31% of students who scored at this level. These results indicate that students tended to shift to either Good or Very Good. Results for Category 3 (Good) also

reflected a decrease, meaning a number of students shifted to Very Good. Results for the Very Good category show a substantial percentage change.

Interpretation/Identification of Facts

	% change
Poor	-49%
Mediocre	-31%
Good	-17%
Very Good	42%

To test for the effect of instruction in the students' critical thinking ability, we used a paired t-test because it has much more statistical power when the difference between groups is small relative to the variation within groups. The hypothesis we are interested in testing is if there is a difference between the means of the two variables (pre and post test results).

We checked for the assumptions underlying the repeated samples t-test.

1. The observations are independent of each other

2. The dependent variable is measured on an interval scale

3. The differences are normally distributed in the population.

Results

Paired Samples Statistics

	Mean	N	Std. Deviation	Std. Error Mean
pre_1A	2.94	724	1.072	.040
post_1A	3.31	724	.950	.035

The table shows that the post-test mean scores are higher.

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	pre_1A & post_1A	724	.352	.000

There is a moderate positive correlation. People who did well on the pre-test also did well on the post-test. The correlation is significant at the 0.05 level.

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	pre_1A - post_1A	-.366	1.155	.043	-.450	-.282	-8.529	723	.000

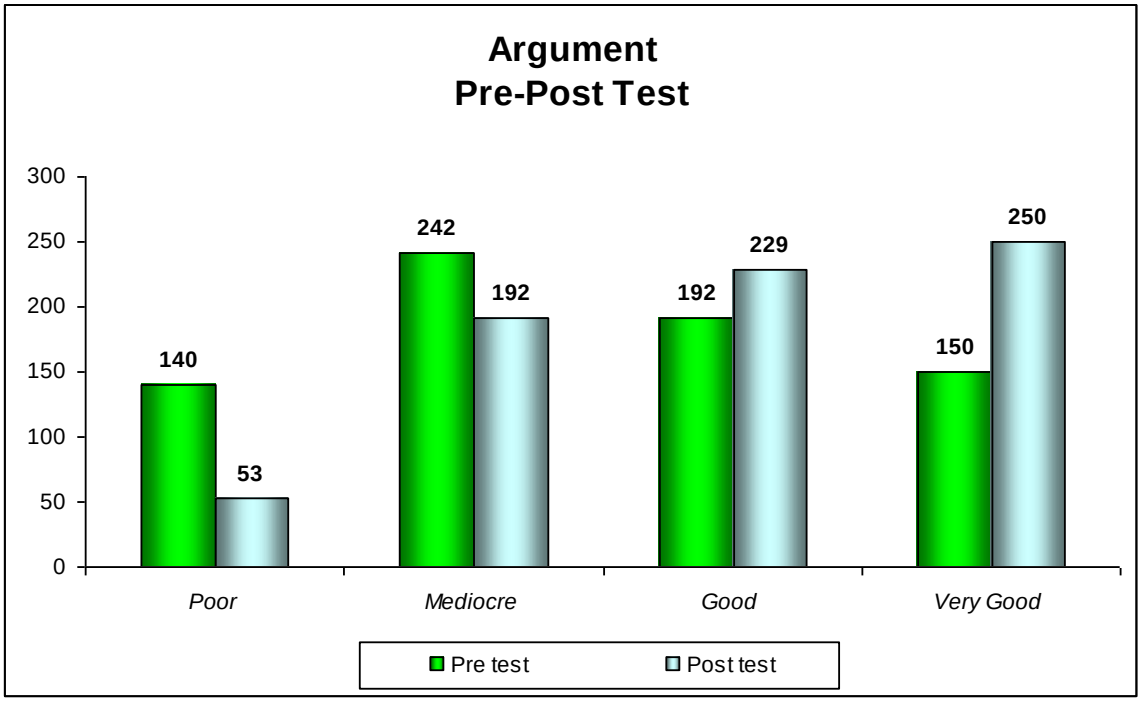
Hypothesis:
Null: There is no significant difference between the means of the two variables (Pre-Post).
Alternate: There is a significant difference between the means of the two variables (Pre-Post).

At the $\alpha = 0.05$ level of significance, the results are statistically significant: a significant increase in the ability of interpretation and identification of facts occurred ($t(723) = -8.529$, $p = .000$). We reject the null hypothesis in favor of the alternative: there is a difference between the means of the two variables (pre and post).

Criterion 2: Argument

	Pre-Test		Post-Test	
	Frequency	Percent	Frequency	Percent
Poor	140	19%	53	7%
Mediocre	242	33%	192	27%
Good	192	27%	229	32%
Very Good	150	21%	250	35%
Total	724	100%	724	100%
Missing	0	0%	0	0%
	724	100%	724	100%

For Criterion 2, the assessment results indicate a shift between the pre and post test for category 1 (Poor). The results show a 62% decrease of students who during the pre-test did poorly to provide a solid argument to the problem. Results for Category 2 (Mediocre) also show a decline of 21% of students who scored at this level. These results indicate that students tended to shift to either Good or Very Good. Results for Category 3 (Good) showed an increase, meaning more students received a grade of Good in this area.. Results for the Very Good category show a substantial percentage change.



Argument

	% change
Poor	-62%
Mediocre	-21%
Good	19%
Very Good	67%

Results

Paired Samples Statistics

	Mean	N	Std. Deviation	Std. Error Mean
pre_2A	2.49	724	1.026	.038
post_2A	2.93	724	.949	.035

The table shows that the post-test mean scores are higher.

Paired Samples Correlations

		N	Correlation	Sig.
Pair 2	pre_2A & post_2A	724	.192	.000

Although there is a low positive correlation, the correlation is significant at the 0.05 level.

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 2	pre_2A - post_2A	-.448	1.256	.047	-.539	-.356	-9.585	723	.000

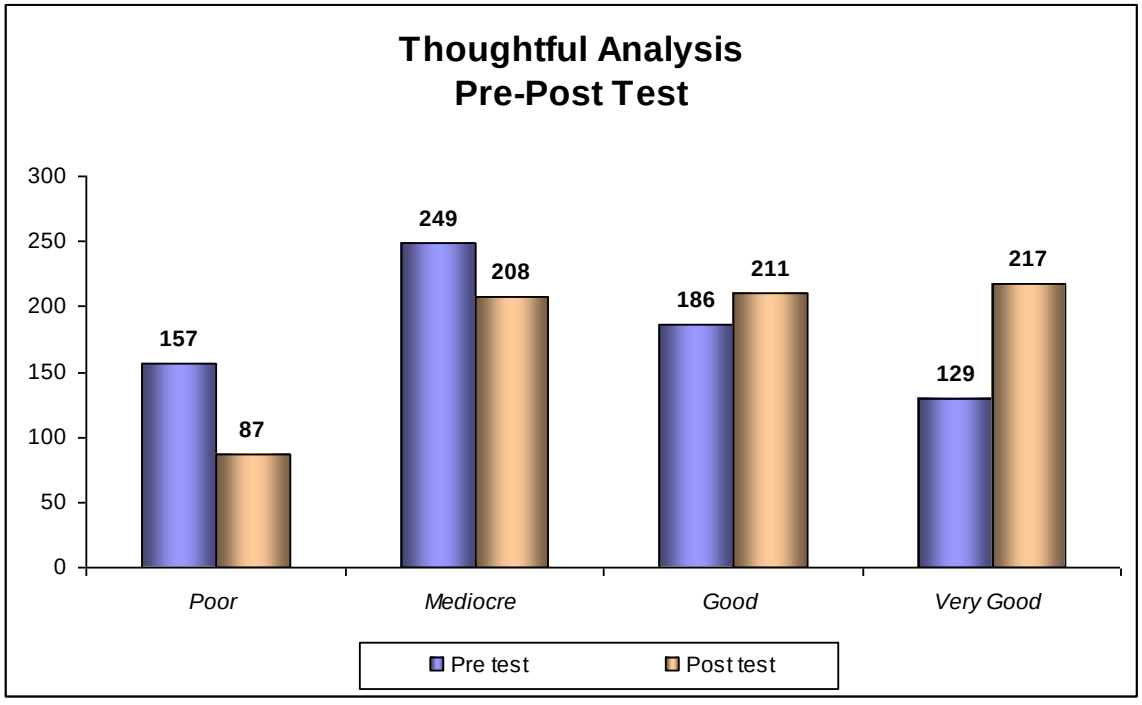
Hypothesis:
Null: There is no significant difference between the means of the two variables (Pre-Post).
Alternate: There is a significant difference between the means of the two variables (Pre-Post).

At the $\alpha = 0.05$ level of significance, the results are statistically significant: a significant increase in the ability of interpretation and identification of facts occurred ($t(723) = -9.585$, $p = .000$). We reject the null hypothesis in favor of the alternative: there is a difference between the means of the two variables (pre and post).

Criterion 3: Thoughtful Analysis

	Pre-Test		Post-Test	
	Frequency	Percent	Frequency	Percent
Poor	157	22%	87	12%
Mediocre	249	34%	208	29%
Good	186	26%	211	29%
Very Good	129	18%	217	30%
Total	721	100%	723	100%
Missing	3	0%	1	0%
	724	100%	724	100%

For Criterion 3, the assessment results indicate a shift between the pre and post test for category 1 (Poor). The results show a 45% decrease of students who during the pre-test did poorly to provide a thoughtful analysis of the problem. Results for Category 2 (Mediocre) also show a decline of 16% of students who scored at this level. These results indicate that students tended to shift to either Good or Very Good. Results for Category 3 (Good) showed an increase of 13%, meaning more students received a grade of Good in this area. Results for the Very Good category show a considerable percentage change (68%).



Thoughtful Analysis

	% change
Poor	-45%
Mediocre	-16%
Good	13%
Very Good	68%

Results

Paired Samples Statistics

	Mean	N	Std. Deviation	Std. Error Mean
pre_3A	2.40	720	1.018	.038
post_3A	2.77	720	1.009	.038

The table shows that the post-test mean scores are higher.

Paired Samples Correlations

		N	Correlation	Sig.
Pair 3	pre_3A & post_3A	720	.282	.000

Although there is a low positive correlation, the correlation is significant at the 0.05 level.

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 3	pre_3A - post_3A	-.375	1.215	.045	-.464	-.286	-8.284	719	.000

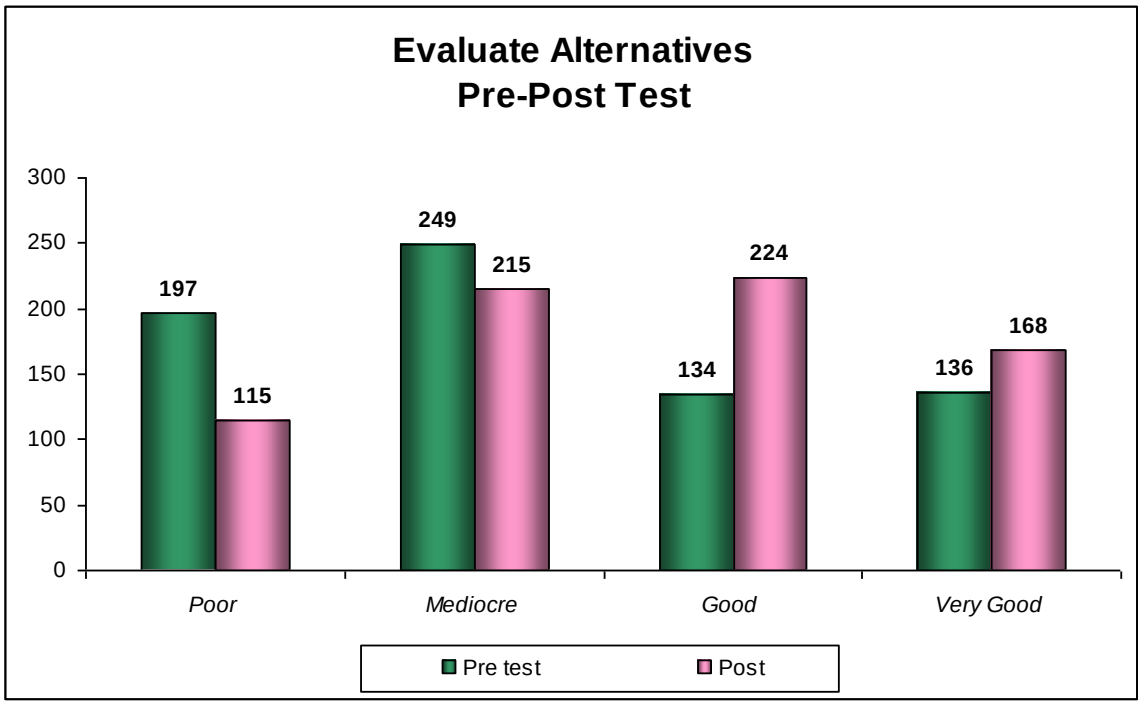
Hypothesis:
Null: There is no significant difference between the means of the two variables (Pre-Post).
Alternate: There is a significant difference between the means of the two variables (Pre-Post).

At the $\alpha = 0.05$ level of significance, the results are statistically significant: a significant increase in the ability of providing a thoughtful analysis occurred ($t(723) = -8.284$, $p = .000$). We reject the null hypothesis in favor of the alternative: there is a difference between the means of the two variables (pre and post).

Criterion 4: Evaluate Alternatives

	Pre-Test		Post-Test	
	Frequency	Percent	Frequency	Percent
Poor	197	27%	115	16%
Mediocre	249	34%	215	30%
Good	134	19%	224	31%
Very Good	136	19%	168	23%
Total	716	99%	722	100%
Missing	8	1%	2	0%
	724	100%	724	100%

For Criterion 4, the assessment results indicate a shift between the pre and post test for category 1 (Poor). The results show a 42% decrease of students who during the pre-test did poorly to evaluate alternatives to the problem. Results for Category 2 (Mediocre) also show a decline of 14% of students who scored at this level. These results indicate that students tended to shift to either Good or Very Good. Results for Category 3 (Good) showed a substantial increase of 67%, meaning more students received a grade of Good in this area. Results for the Very Good category show a percentage change (24%), although not as considerable as in the previous sections.



Evaluate Alternatives

	% change
Poor	-42%
Mediocre	-14%
Good	67%
Very Good	24%

Results

Paired Samples Statistics

	Mean	N	Std. Deviation	Std. Error Mean
pre_4A	2.29	714	1.066	.040
post_4A	2.61	714	1.012	.038

The table shows that the post-test mean scores are higher.

Paired Samples Correlations

		N	Correlation	Sig.
Pair 4	pre_4A & post_4A	714	.229	.000

Although there is a low positive correlation, the correlation is significant at the 0.05 level.

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 4	pre_4A - post_4A	-.324	1.291	.048	-.418	-.229	-6.697	713	.000

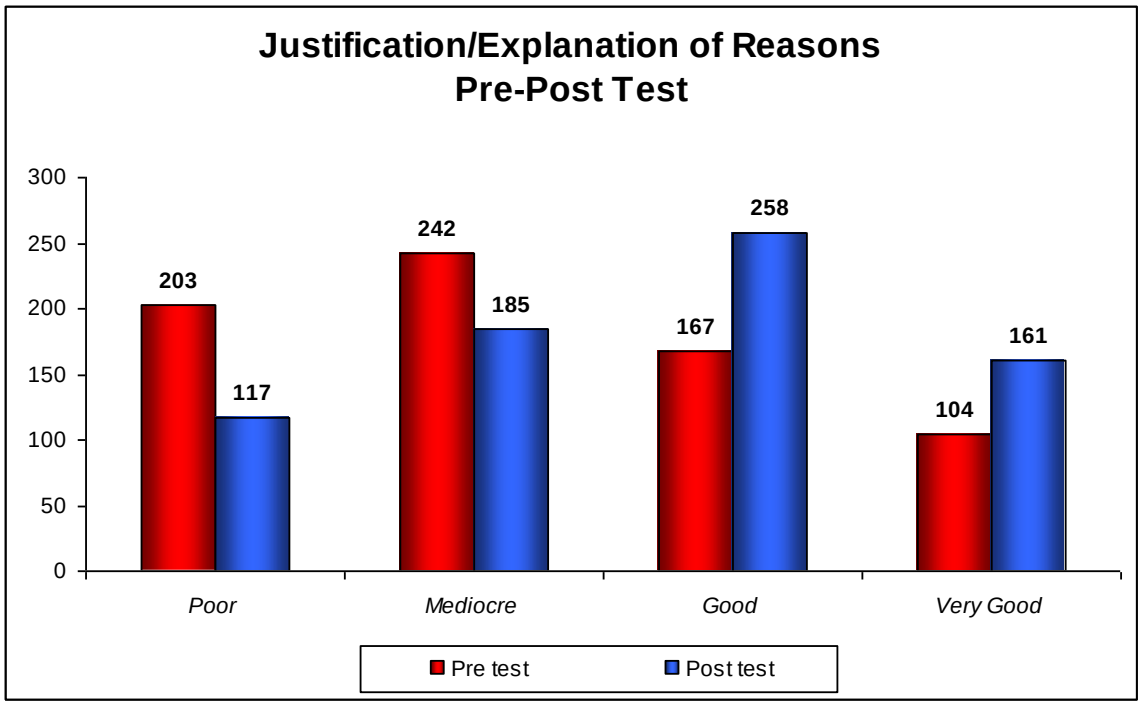
Hypothesis:
Null: There is no significant difference between the means of the two variables (Pre-Post).
Alternate: There is a significant difference between the means of the two variables (Pre-Post).

At the $\alpha = 0.05$ level of significance, the results are statistically significant: a significant increase in the ability of providing a thoughtful analysis occurred ($t(713) = -6.697$, $p = .000$). We reject the null hypothesis in favor of the alternative: there is a difference between the means of the two variables (pre and post).

Criterion 5: Justification/Explanation of Reasons

	Rater A		Rater B	
	Frequency	Percent	Frequency	Percent
Poor	203	28%	117	16%
Mediocre	242	33%	185	26%
Good	167	23%	258	36%
Very Good	104	14%	161	22%
Total	716	99%	721	100%
Missing	8	1%	3	0%
	724	100%	724	100%

For Criterion 5, the assessment results indicate a shift between the pre and post test for category 1 (Poor). The results show a 42% decrease of students who during the pre-test did poorly to provide justifications/explanation to their reasons. Results for Category 2 (Mediocre) also show a decline of 24% of students who scored at this level. These results indicate that students tended to shift to either Good or Very Good. Results for Category 3 (Good) showed a substantial increase of 54%, meaning more students received a grade of Good in this area. Results for the Very Good category also show an impressive percentage change (55%).



Justification/Explanation of Reasons

	% change
Poor	-42%
Mediocre	-24%
Good	54%
Very Good	55%

Results

Paired Samples Statistics

	Mean	N	Std. Deviation	Std. Error Mean
pre_5A	2.24	714	1.019	.038
post_5A	2.63	714	1.000	.037

The table shows that the post-test mean scores are higher.

Paired Samples Correlations

		N	Correlation	Sig.
Pair 5	pre_5A & post_5A	714	.306	.000

Although there is a low positive correlation, the correlation is significant at the 0.05 level.

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 5	pre_5A - post_5A	-.398	1.190	.045	-.485	-.310	-8.934	713	.000

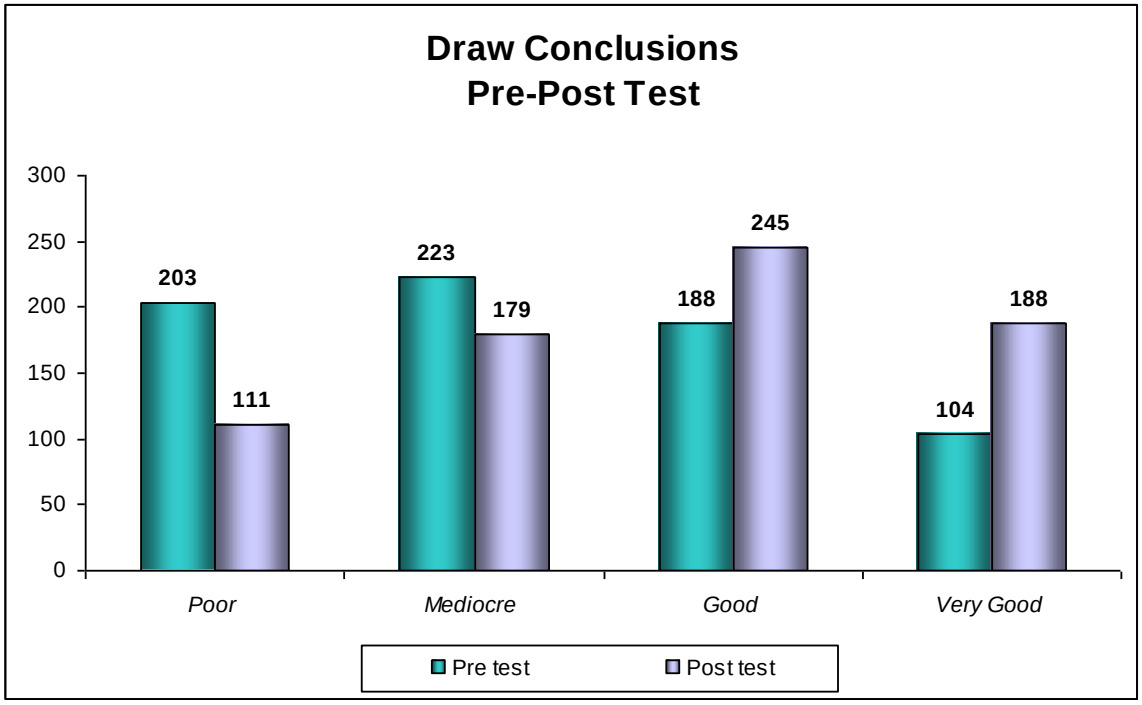
Hypothesis:
Null: There is no significant difference between the means of the two variables (Pre-Post).
Alternate: There is a significant difference between the means of the two variables (Pre-Post).

At the $\alpha = 0.05$ level of significance, the results are statistically significant: a significant increase in the ability of providing a thoughtful analysis occurred ($t(723) = -8.934$, $p = .000$). We reject the null hypothesis in favor of the alternative: there is a difference between the means of the two variables (pre and post).

Criterion 6: Draw Conclusions

	Pre-Test		Post-Test	
	Frequency	Percent	Frequency	Percent
Poor	203	28%	111	15%
Mediocre	223	31%	179	25%
Good	188	26%	245	34%
Very Good	104	14%	188	26%
Total	718	99%	723	100%
Missing	6	1%	1	0%
	724	100%	724	100%

For Criterion 6, the assessment results indicate a shift between the pre and post test for category 1 (Poor). The results show a 45% decrease of students who during the pre-test did poorly to draw conclusions. Results for Category 2 (Mediocre) also show a decline of 20% of students who scored at this level. These results indicate that students tended to shift to either Good or Very Good. Results for Category 3 (Good) showed an increase of 30%, meaning more students received a grade of Good in this area. Results for the Very Good category also show an remarkable percentage change (81%).



Draw Conclusions

	% change
Poor	-45%
Mediocre	-20%
Good	30%
Very Good	81%

Results

Paired Samples Statistics

	Mean	N	Std. Deviation	Std. Error Mean
pre_6A	2.27	717	1.026	.038
post_6A	2.70	717	1.019	.038

The table shows that the post-test mean scores are higher.

Paired Samples Correlations

		N	Correlation	Sig.
Pair 6	pre_6A & post_6A	717	.288	.000

Although there is a low positive correlation, the correlation is significant at the 0.05 level.

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 6	pre_6A - post_6A	-.432	1.220	.046	-.522	-.343	-9.488	716	.000

Hypothesis:
Null: There is no significant difference between the means of the two variables (Pre-Post).
Alternate: There is a significant difference between the means of the two variables (Pre-Post).

At the $\alpha = 0.05$ level of significance, the results are statistically significant: a significant increase in the ability of providing a thoughtful analysis occurred ($t(723) = -9.488$, $p = .000$). We reject the null hypothesis in favor of the alternative: there is a difference between the means of the two variables (pre and post).

Conclusions

According to the assessment results, students performed exceptional well during the post test. The findings show that the results are statistically significant in all areas.

As we continue to review and analyze the results, several questions will come to mind regarding the characteristics of a pre-post test. Few tests exist that are perfectly valid, can be generalized to all populations, and assess students in all of the areas of interest. However, there are some critical features of pre-post tests that should not be ignored. The first question to consider has to do with the instruments used during this assessment. For example, how reliable was each test? What would be the extent to which a student would be expected to perform similarly across multiple administrations of the test under similar conditions? Another question to ponder is, how valid was each test? To what extent the test is measuring what it is expected to measure?

As the analysis continues, we will have to review the construct validity for each test. Do the results provide evidence that each test was measuring the content and skills (or "construct", in this case critical thinking) that it claims to measure? Do we have enough evidence that the tasks on the tests adequately covered the area of interest? In the test prepared by each department, we will need to analyze whether or not the tasks on the instrument were aligned with the skills associated with critical thinking. For example, did all the items represent the full range of skills associated with critical thinking? In general, we will be looking at how skills not associated with this content area, such as knowledge of science or social studies, influenced performance on the test.

Final steps

Once all the necessary computations are done, the results will be analyzed and reviewed by faculty. Through several meetings of the Assessment Committee, faculty will discuss the results of the data collected and how they will use the results to improve student learning. Some examples of action steps are a) changing the way some material in a course is presented by the instructor, b) adapting a course to include newly developed technologies, and c) adding a new course to the program. This has been an exciting study that will lead the process to implement other assessment projects at the College.

Appendix A
Rubric

Criteria	**Very Good (4)**	**Good (3)**	**Mediocre (2)**	**Poor (1)**
Interpretation/ Identification of facts				
Argument				
Thoughtful Analysis				
Evaluate Alternatives				
Justification/Explanation of Reasons				
Draw Conclusions				