Pilot Assessment of Critical Thinking - Spring 2010
Summary Report

Purpose of the Pilot Assessment:
The purpose for the pilot assessment of critical thinking was to provide a trial for the assessment process and the critical thinking rubric. Student samples were made available through two instructors who willingly committed to participate in this process. Students were not representative of the general student body, and none of them, to the best or our knowledge, were graduating seniors, or even had senior status. There was no expectation that findings from the assessment would provide reliable information about critical thinking abilities in graduating seniors at our institution.

Process:
In the spring of 2010, the Higher Learning Commission Assessment Academy (HLC AA) team developed a rubric for assessing critical thinking. Two instructors on our campus volunteered to let the team use samples of their students' work to assess the level of critical thinking demonstrated by their students. Student products were evaluated by pairs of readers, using the critical thinking rubric. Scores were recorded, and reliability statistics were applied to the raw scores.

The first group of student papers was evaluated by members of the Outcomes Assessment Committee 1 on May 3, 2010. There were seven readers total, and after an initial norming session, each subsequent paper was read by two evaluators who independently scored each paper using the critical thinking rubric. Scores on each component of the rubric were recorded, as well as composite scores for each paper.

The second group of student papers was evaluated by members of the HLC AA team, on August 4, 2010. There were three readers total, and after an initial norming session, each subsequent paper was read by two evaluators who independently scored each paper using the critical thinking rubric. Scores on each component of the rubric were recorded, as well as composite scores for each paper.

In both cases, two student products were used in the norming process. All participants read and rated the first paper individually. Then ratings were compared and discussed until consensus was reached. After the second paper was completed in the same manner, the both groups felt they were ready to move forward to evaluating all papers.

In the first case, student products were draft papers wherein students reflected on the research process. In the second case student products were also draft papers, but the content of those papers was an argument presenting two or more opposing views to an issue.

In both cases, the following evaluative process was used:

- Two copies of each paper were printed and numbered.
- Each reader was assigned a 'reader number'.
- Scoring rubrics (attached) were provided with each paper, and scores ranged from 1-4 (1=beginning, 2=developing, 3=competent, 4=accomplished) on 5 criteria (analyzing information, drawing conclusions, presenting multiple solutions/positions, synthesizing ideas, and identifying salient arguments within own worldview).
- Paper number and reader number were identified on each scoring rubric.
- Each paper was read and scored by two independent readers.
- Scores were entered into an excel worksheet (attached).

Findings:
Evaluator scores were transferred to SPSS. Krippendorf's alpha, a separate statistical test for Inter-Rater Reliability (IRR) was run on each of the separate dimensions of the rubric. Krippendorf's alpha was used because:

- It is preferred for content analysis (rendering judgments of text-based information)
- It handles small data sets
- It accommodates missing data
- It uses bootstrapping for more precise measurement

Krippendorf's alpha measures agreement of evaluators who rate a set of items into distinct and mutually exclusive categories. The observed disagreement between evaluators is corrected by the amount of disagreement expected by chance. (Poesio and Artstein 2005)

The value of alpha can range from -1.0 (complete disagreement) to 0 (unreliability of measurement) to 1.0 (complete agreement). The higher alpha is in the positive direction, the greater level of agreement between evaluators.

In general measures of agreement, the following guidelines are given:

0.8 and 1	Very good agreement
0.6 and 0.79	Good agreement
0.4 and 0.59	Moderate agreement
0.2 and 0.39	Fair agreement
0.0 and 0.19	Poor agreement

Results from test application of Critical Thinking Analysis Rubric (May 2010)*

Rubric Dimension	Krippendorf's α	Interpretation
Analyzing Information	0.28	Fair agreement
Drawing well-supported conclusions	-0.06	unreliable measure
Presenting multiple solutions, positions or perspectives	0.43	Moderate agreement
Synthesizing ideas into a coherent whole	-0.13	unreliable measure
Identifying arguments while forming and situating own worldview in a larger context	0.29	Fair agreement

Results from test application of Critical Thinking Analysis Rubric (August 2010)*

Rubric Dimension	Krippendorf's α	Interpretation
Analyzing Information	0.60	Good agreement
Drawing well-supported conclusions	0.34	Fair agreement
Presenting multiple solutions, positions or perspectives	0.74	Good agreement
Synthesizing ideas into a coherent whole	0.54	Moderate agreement
Identifying arguments while forming and situating own worldview in a larger context	0.51	Moderate agreement

**Analysis by Patricia MacGregor-Mendoza*

Results from the IRR statistics indicate that the evaluators' interpretation/application of the rubric on the first group of student papers was inconsistent and did not render a satisfactory level of agreement on any of the rubric dimensions. Application of the rubric on the second group of student papers revealed greater consistency, and reached levels of 'good agreement' on two of the 5 dimensions.

Conclusion:
As anticipated, it is clear that data from the pilot assessment of critical thinking is insufficient to determine the level of ability of students at NMSU to think critically.

Discussion:
The pilot assessment of critical thinking was successful in respect to the purpose of the assessment: To provide a trial for the assessment process and the critical thinking rubric. In many respects, the process for collecting and evaluating student products worked well. What did not work as effectively was the consistent application of the rubric to the student product. Evaluators indicated that the rubric was sometimes difficult to apply to student work, because there was ambiguity in the minds of the evaluators about particular aspects of the rubric, as well as ambiguity in student work. Additionally, evaluators were not clear on the specific assignment students were given in the first group of papers, and the assignment seemed to be more ambiguous than that of the second group of student papers. Also, there seemed to be more correlation with the rubric and the assignment in the second group of papers than in the first. Finally, there were fewer evaluators for the second group of papers, and they were more likely to discuss student products among themselves prior to rating each paper.

Suggestions were made by both groups of evaluators to make adjustments to the rubric. Additionally, both groups indicated it would be advantageous to have copies of the assignment, complete with any specific instructions, that was given to the students.

It is also clear that training and norming of evaluators must be improved.

Specific Suggestions to improve IRR results:

- Examine instructions provided to students/instructor regarding the preparation/evaluation of their writing samples.
- Examine the rubric dimensions and descriptions for each category for clarity, mutual exclusivity.
- Examine training provided to evaluators.

Poesio, M. & Artstein, R. (2005). *The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account.,* Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the sky, pp. 76-83, Ann Arbor: Association for Computational Linguistics.
August 4, 2010 Data:

Data from Group 1 – May, 2010

Paper #	Reader #	Analy Info	Conclusions	Multi Sol	Synthesis	Worldview	Composite
T3	5	3	3	3	3	3	3
2	7	3	3	3	2		2.75
2	3	4	3	3	3	2	3
3	5	3	3	4	2	2	2.8
3	4	3	2	2	2	2	2.2
1	2	1	2	2	1		1.5
1	1	1	1	2	2	2	1.6
4	3	4	3	3	3	3	3.2
4	7	4	3	4	4	3	3.6
7	3	2	2	2	2		2
6	7	3	3	3	3		3
5	4	3	3	3	3	2	2.8
8	3	2	2	2	2		2
7	1	2	3	3	2	2	2.4
5	5	4	4	3	4	3	3.6
8	7	2	3	2	3		2.5
15	3	3	2	2	2		2.25
9	1	3	3	3	3	2	2.8
10	5	2	2	2	2	2	2
10	7	4	4	4	4	3	3.8
12	7	3	2	2	2	2	2.2
11	1	3	3	3	2	2	2.6
11	5	4	3	3	3		3.25
12	1	1	2	1	2	2	1.6
13	7	4	4	3	3	3	3.4
14	3	3	2	2	2		2.25
13	5	3	3	3	4	3	3.2
6	6	2	2	3	2	1	2
16	7	4	4	4	4	4	4
16	1	4	3	4	3	3	3.4
T3	7	2	2	3	3	3	2.6
9	4	3	2	2	2	2	2.2
14	4	3	3	3	3	2	2.8
15	4	2	2	1	2	2	1.8
T2	2	3	3	3	4		3.25
T2	1	3	3	3	3	2	2.8
T1	3	2	2	2	2		2
T1	5	2	2	2	1	2	1.8
AVG		**2.815789474**	**2.657894737**	**2.68421053**	**2.60526316**	**2.37037037**	**2.630263158**

Data from Group 2 – August, 2010

Paper #	Reader #	Analy Info	Conclusions	Multi Sol	Synthesis	Worldview	Composite
10	1	2	1	2	1	1	1.4
10	2	2	1	2	1	1	1.4
10	3	2	1	2	1	1	1.4
12	1	2	1	2	1	2	1.6
12	2	2	1	2	2	1	1.6
11	2	1	1	1	1	1	1
11	3	1	1	1	1	1	1
9	1	2	3	3	2	3	2.6
9	3	3	3	3	2	3	2.8
8	1	1	2	2	2	2	1.8
8	2	1	1	2	2	2	1.6
7	1	2	2	1	2	2	1.8
7	3	2	2	2	2	1	1.8
6	2	2	2	2	1	2	1.8
6	3	2	2	2	2	2	2
5	1	1	1	2	2	1	1.4
5	2	2	2	2	2	2	2
4	1	1	2	1	1	2	1.4
4	3	1	1	1	1	2	1.2
3	1	2	2	2	2	1	1.8
3	2	2	1	2	2	1	1.6
2	1	3	2	3	2	3	2.6
2	3	3	2	3	2	3	2.6
1	2	1	1	1	1	2	1.2
AVG		**1.791666667**	**1.583333333**	**1.91666667**	**1.58333333**	**1.75**	**1.725**