



Office of Institutional Research

To: Scholarship Action Group, ESIMT &
Curriculum Committee
CC: Critical Thinking Instructors
From: Darby Hiller
Date: September 10, 2007
Subject: Critical Thinking Artifact Analysis (Spring 2007)

Executive Summary

NMC measures student learning on the general education outcomes by assessing original student work. The student work (artifacts) is generated through the students' own coursework and offers an authentic assessment mechanism. In May 2007, faculty and staff gathered over two scoring days to score student work measuring critical thinking. NMC's general education goal is that near-graduating students (those with 52 or more college credits) will perform minimally at the sufficient level on the [outcome rubric](#).

Key findings:

- 72.4% performed at the sufficient level or above (an 8% drop from the previous critical thinking scoring cycle)
- Strongest skill: Identifying the issue or problem
- Weakest skill: Demonstrating an understanding of different perspectives
- Many un-scorable artifacts contributed to a 43% attrition rate and a small sample size
- 94% inter-reader reliability rate was consistent with previous years

Methodology and Inter-reader Reliability

In spring 2007, the population of near-graduates was 1126. The sample was drawn on January 25 at the end of the registration drop/add period to reduce attrition rate for students still deciding on a schedule of courses, and to give instructors selected to submit artifacts the maximum advanced notice. The sampling method used for artifacts targeted students.

First, the population of students was determined. From the population the requisite sample size randomly drawn was 287 for a +/- 5% margin of error. To further account for attrition, 32% more students were added to the sample, for 379 possible artifacts. Second, the courses directly supporting critical thinking were selected in the order of near graduate enrollment. On January 26, instructors of the selected courses were notified in writing by the Vice President for Educational Services that they had been selected to submit artifacts for the

designated list of students. Each instructor was also sent a copy of the Artifact Guidelines (see appendix A).

Scoring took place on Tuesday, May 8th, and Thursday, May 9th with the assistance of 31 faculty and staff. Scorers applied the critical thinking rubric to the artifacts and assigned a score for each capability. To optimize reliability, two scorers read each piece of student work and assigned a score from zero (deficient) to three (proficient) for each of the five capabilities on the rubric. If the sum of the capability scores from one reader differed by three or more points (on a 15-point scale) from the second reader, a third reader was required to score the artifact.

A difference of three or more points was chosen because three points represents a 20% difference in opinion between readers, which we define as a significant difference. The final score for an artifact was the average score between the two readers, unless a third reader was needed. If there was a third reader (thus, three sets of scores), the final score for the artifact was the average of the two closest scores within a 3-point difference. Artifacts with total scores differing by three or more after the third reader were dropped from the outcome performance analysis. By this means, *inter-reader reliability* is assured statistically. The scorers also took steps to ensure inter-reader reliability qualitatively as well. Prior to scoring of artifacts, scorers independently read a set of student work that reflected a range of assignments. The scorers came together for a norming session to review their scores, to reconcile inconsistent patterns, and to develop a shared understanding of the rubrics.

For the first and second read of the scorable artifacts (N=231), 65% were within acceptable reliability limits (Table 1). Third readers were required to score 81 artifacts (35%).

Difference	Frequency	Percent	Cumulative Percent
0	42	18.2	18.2
1	61	26.4	44.6
2 Acceptable Reliability	47	20.3	64.9
3	37	16.0	81.0
4	18	7.8	88.7
5	10	4.3	93.1
6	8	3.5	96.5
7	5	2.2	98.7
8	1	.4	99.1
9	1	.4	99.6
11	1	.4	100.0
Total	231	100.0	

When considering the scores of the third reader, the reliability rate increased to 93.9% (Table 2). There were 14 artifacts that were outside the acceptable reliability limits and as such were thrown out of the analysis. The final outcomes analysis included 217 artifacts (reasons for the small sample size provided in the last section of the report; see Table 6 for a list of courses and number of artifacts included).

Difference	Frequency	Percent	Cumulative Percent
0	62	26.8	26.8
1	91	39.4	66.2
2 Acceptable Reliability	64	27.7	93.9
3	11	4.8	98.7
4	3	1.3	100.0
Total	231	100.0	

NMC's goal is to increase the reliability rate through training, reader norming sessions, and rubric improvements. Our goal is to have agreement within one point a majority of the time and to have scores not more than two points apart 100% of the time. Table 3 shows the trends in inter-reader reliability for critical thinking artifacts over the semesters in which it was tracked.

	2004	2005	2007
Number of artifacts scored	50	143	231
First two reads	64%	67%	65%
Addition of third read	94%	95%	94%
Number of artifacts thrown out of analysis	3	8	14

A sample of 217 artifacts yields a margin of error of +/- 6%. The results cannot be generalized to all of our near graduates with confidence. However, the descriptive statistics for these specific scores can be useful in guiding our assessment improvement strategies, and in confirming trends.

Results

Overall, 72.4% of the near-graduates scored sufficient or better on the Critical Thinking rubric (Table 4, Valid Percent column). In spring 2005, 80.0% scored sufficient or better

(N=135). On average, near-graduates performed sufficiently on all five capabilities (Table 5).

Level of performance is defined on the critical thinking rubric as:

- Deficient – mean score 0.0 to 0.7
- Developing – mean score 0.8 to 1.5
- Sufficient – mean score 1.6 to 2.3
- Proficient – mean score 2.4 to 3.0

Spring 2007	Frequency	Valid Percent	Cumulative Percent
Deficient	9	4.1	4.1
Developing	51	23.5	27.6
Sufficient	87	40.1	67.7
Proficient	70	32.3	100.0
Total	217	100.0	

	Spring 2007				Spring 2005	
	Score	N	Mean	Std. Dev.	N	Mean
ct1: Identifies issue or problem	Sufficient	214	2.21	.70	121	2.43
ct2: Demonstrates an understanding of different perspectives	Sufficient	139	1.72	.80	65	1.78
ct3: Uses information to results issue or problem	Sufficient	211	2.07	.76	135	2.23
ct4: Applies reasoning to resolve issue or problem	Sufficient	206	1.85	.80	135	2.08
ct5: Draws conclusions that resolves issue or problem	Sufficient	211	1.77	.80	135	1.89
Overall Score	Sufficient	217	1.96	.70	135	2.13

The strongest skill was CT1: identifies the issue or problems (mean = 2.21). The weakest skill was CT2: demonstrates an understanding of different perspectives (mean = 1.72). This is consistent with the spring 2005 artifact findings. Overall students did not perform as well in 2007 as they did in 2005.

What this tells us - Check

The methodology used to analyze progress on the Critical Thinking outcome in spring 2007 allowed us to directly measure the NMC goal that near-graduates perform sufficient or better on the general education outcome. The large sample of artifacts gives us greater confidence that the results reflect the skills of our students.

Adjusting the Curriculum

Overall the percentage of near-graduates that performed sufficiently or better on the critical thinking outcome decreased by nearly 8% from 80% in spring 2005 to 72% in spring 2007. As far as the Scholarship Action Group knows, there were no formal curricular changes after the assessment cycle in 2005-2006 meant to target improved student learning in critical thinking. These results are being reported to ESIMT and Curriculum Committee so that adjustments and plans for improvement can take place this year.

The ability to demonstrate an understanding of, or at least a consideration of, different perspectives in drawing conclusions (capability #2) is the weakest skill of our students and has been through three assessment cycles. This finding is consistent whether the artifact method or the CAAP critical thinking test is used to measure student learning.

Adjusting the Assessment Process

The attrition rate of scorable artifacts was 43%, which exceeded even the conservative estimate of 32% prior to generating the sample. Attrition of artifacts occurred in four ways:

- 1) 76 (20%) of the requested 379 artifacts were not handed in to be scored, this may be due to students dropping the course or not handing in the assigned piece of work
- 2) 37 artifacts (10%) from six different assignments were deemed un-scorable by the faculty scorers on scoring day because they did not adequately measure the critical thinking capabilities for whatever reason
- 3) 35 artifacts (10%) from four multiple choice assignments were thrown out of the analysis because multiple choice questions and answers do not provide evidence of critical thinking unless each question is aligned with one and only one capability it is meant to assess
- 4) 14 artifacts (3%) scored by a third reader were outside the acceptable inter-reader reliability limits and as such were thrown out of the analysis

Of considerable benefit to the artifact scoring process in spring 2006 was the introduction of the [Artifact Guidelines](#). Instructors submitting Communications artifacts followed the guidelines. This allowed for increased efficiency in the scoring process and minimized attrition. With the critical thinking artifacts in spring 2007, it is apparent from the assignments that were submitted that less attention was paid to the Artifact Guidelines. Unfortunately, this

meant that on scoring day there were many assignments and sets of artifacts that were deemed unscorable. This poses a problem on at least three levels.

First, there are too few scored artifacts from the near-graduates to be able to generalize to the population, making it difficult to measure our general education goals. Second, it is apparent that instructors may not have assignments that lead students to demonstrate critical thinking skills, thus making it difficult to assess learning in this area. Third, the actual scoring of the artifacts becomes less efficient as the variation in assignment and artifact format increases.

In the future, the Scholarship Action Group will need to improve the education of and communication with faculty in general, and instructors submitting artifacts specifically, as to the appropriateness of the assignments used to generate the artifacts for the purpose of general education assessment. There are several [model assignments](#) available on the IR website that provide explanation for eliciting critical thinking from students and getting the students to demonstrate their thinking in their work. With the help of the academic area chairs, the Scholarship Action Group will continue to improve our general education assessment methods.

Address questions or comments about this analysis to [Darby Hiller](#), in the Office of Institutional Research. For information about the Scholarship Action Group, contact [Darby Hiller](#) or [Tom Gordon](#), co-chairs.

Table 6. Courses and Number of Scorable Artifacts	Frequency	Percent	Cumulative Percent
AVG 381	4	1.7	1.7
BIO 100	10	4.3	6.1
BIO 105	4	1.7	7.8
BIO 106	4	1.7	9.5
BIO 106L	3	1.3	10.8
BIO 208	4	1.7	12.6
BIO 228	7	3.0	15.6
BUS 150	7	3.0	18.6
BUS 261	4	1.7	20.3
BUS 262	4	1.7	22.1
COM 111	3	1.3	23.4
ECO 201	6	2.6	26.0
ECO 202	14	6.1	32.0
ENG 211	6	2.6	34.6
ENG 271	4	1.7	36.4
HNR 101	16	6.9	43.3
HNR 125	16	6.9	50.2
MGT 241	3	1.3	51.5
MGT 255	4	1.7	53.2
MKT 201	13	5.6	58.9
MTH 106	4	1.7	60.6
MTH 111	9	3.9	64.5
MTH 116	2	.9	65.4
MTH 121	3	1.3	66.7
MTH 122	4	1.7	68.4
MTH 131	5	2.2	70.6
PAR 101	6	2.6	73.2
PHL 202	8	3.5	76.6
PHY 105	8	3.5	80.1
PHY 122	10	4.3	84.4
PHY 222R	8	3.5	87.9
PLS 211	8	3.5	91.3
PSY 211	7	3.0	94.4
PSY 231	4	1.7	96.1
SOC 201	3	1.3	97.4
VCA 225	6	2.6	100.0
Total	231	100.0	

Appendix A

Artifact Guidelines

General Education Outcome Artifact Guidelines

The purpose of the following artifact guidelines is to assist faculty in selecting student work that represents our students' level of performance on NMC's General Education Outcomes. These guidelines are also intended to help the Scholarship Action Group (SAG) and the faculty volunteers score the artifacts. The primary impetus for developing these guidelines came from the faculty volunteers who scored student work in May 2005.

Faculty members scoring artifacts identified the need to have some guidelines for receiving scorable artifacts. A few of these comments are listed below.

- "need a better process for getting scoreable assignments"
- "instructors need more instruction on creating scoreable artifacts"
- "to some extent, I found myself disappointed with the assignment"
- "excellent quality overall with a few minority cases where work is needed in assignment construction"
- "I thought some of these [artifacts] were very difficult to grade"

In a combined effort from the SAG and the Educational Services Instructional Management Team (ESIMT), the following artifact guidelines have been established:

Scorable artifacts represent individual student work.

Group artifacts have not worked in the past because scorers are not able to attribute performance on the outcome capabilities to an individual student. It is critical to the process of assessment to have individual student work because the population to which we intend to generalize the results is made up of individual students. It is fine for the overall project to culminate in group work, but unless an individual's work can be singled out and submitted as an artifact, the group artifact cannot be scored.

Scorable artifacts represent student performance on the general education outcome as defined by the capabilities on the scoring rubric.

Assignments that do not ask the student to demonstrate the outcome capabilities provide significant challenges to those scoring the student work. To help facilitate the scoring of the artifacts, it is extremely helpful when the faculty members submitting the artifacts provide a copy of the assignment that generated the work, AND an explanation of how and where the student was expected to demonstrate the outcome capability. If faculty need assistance in creating assignments that result in scorable artifacts, the SAG, Writing Center, and the academic area chairs are prepared to work with them. If after review of the artifacts and assignment, there

is still ambiguity as to how the student work is to be scored, the SAG may request guidance from the submitting faculty member.

Scorable artifacts represent student work that is part of the graded work for the course.

The assignment that generates the student work to be scored for general education assessment also generates student work that is graded by faculty members for a student's grade in the course. In this way, the artifacts are considered course-embedded and provide a more authentic assessment of a student's level of achievement. Moreover, course-embedded artifacts ensure that students are motivated to demonstrate their best work.

Scorable artifacts can be scored in a reasonable amount of time and meet the above guidelines.

During the scoring of artifacts experience has shown that each faculty volunteer scores an average of 30 artifacts. Lengthy artifacts (large capstone projects) require an extraordinary amount of time to score. At the same time, artifacts that are too brief do not provide enough evidence with which to evaluate the student's performance on the outcome. To streamline the scoring process, an ideal artifact length is 2-10 pages if written work is submitted.

Scorable artifacts are written and in paper format.

For the time being, artifacts for scoring need to be submitted in hard paper copy. As the process of scoring is refined, the possibility of scoring videos of performances, speeches, etc., will be revisited.