



Northwestern  
Michigan  
College

## Office of Institutional Research

To: Scholarship Action Group  
From: Darby Hiller  
Date: January 29, 2005  
Subject: Inter-reader Reliability Study (Fall 2004)

On January 12<sup>th</sup>, 2005, ten scorers came together to score 99 pieces of student work. Fifty artifacts represented performance on the Critical Thinking outcome and 49 represented performance on the Communications outcome. For the actual scoring process, 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. It also represents a more rigorous decision point than suggested in current literature, which tends to be a 25% 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 all the scores that were within one standard deviation of the mean. Any artifacts with differences greater than three between the scores of the third reader and the other reader with the standard deviation were thrown out of the outcome performance analysis. By this means, statistically, *inter-reader reliability* is assured. The scorers also took steps to ensure inter-reader reliability qualitatively as well. Prior to the actual scoring of artifacts, scorers independently read a set of student work that reflected a range of assignments. When the scorers came together they reviewed their scores to reconcile inconsistent patterns.

This study shows the extent to which the first two scorers agreed and the number of times a third scorer was required with reliability rate.

**Communications Artifacts.**

A total of 49 communications artifacts were scored. About 76% of the time the readers' scores were within two points (Table 1) on a 15-point scale. NMC's goal is to increase that reliability rate through training and reader norming sessions. Our goal is to have exact agreement a majority of the time and to have scores not more than two points apart 100% of the time.

	Frequency	Percent	Valid Percent	Cumulative Percent
0	12	24.5	24.5	24.5
1	14	28.6	28.6	53.1
2	11	22.4	22.4	75.5
3	6	12.2	12.2	87.8
4	4	8.2	8.2	95.9
5	2	4.1	4.1	100.0
Total	49	100.0	100.0	

There were 12 (24%) communications artifacts that needed a third reader. When factoring scores of the third reader the reliability rate increased to 100% (Table 2). All the artifacts were used in the outcome performance analysis.

	Frequency	Percent	Valid Percent	Cumulative Percent
0	19	38.8	38.8	38.8
1	19	38.8	38.8	77.6
2	11	22.4	22.4	100.0
Total	49	100.0	100.0	

**Critical Thinking Artifacts.**

A total of 50 critical thinking artifacts were scored. Sixty-four percent of the time the readers' scores were within two points (Table 3). NMC's goal is to increase that reliability rate through training and reader norming sessions. Our goal is to have exact agreement a majority of the time and to have scores not more than two points apart 100% of the time.

	Frequency	Percent	Valid Percent	Cumulative Percent
0	9	18.0	18.0	18.0
1	10	20.0	20.0	38.0
2	13	26.0	26.0	64.0
3	5	10.0	10.0	74.0
4	3	6.0	6.0	80.0
5	2	4.0	4.0	84.0
6	3	6.0	6.0	90.0
7	4	8.0	8.0	98.0
8	1	2.0	2.0	100.0
Total	50	100.0	100.0	

There were 18 (36%) critical thinking artifacts that needed a third reader. With the addition of the third reader, the reliability rate increased to 94% (Table 4). The three scores that were outside the acceptable reliability limit will be dropped from the outcome performance analysis.

	Frequency	Percent	Valid Percent	Cumulative Percent
.00	14	28.0	28.0	28.0
1.00	14	28.0	28.0	56.0
2.00	19	38.0	38.0	94.0
3.00	2	4.0	4.0	98.0
6.00	1	2.0	2.0	100.0
Total	50	100.0	100.0	