



Office of Institutional Research

To: Scholarship Action Group
CC: Curriculum Committee & ESIMT
From: Darby Hiller
Date: May 19, 2005
Subject: Inter-reader Reliability Study (Spring 2005)

In May, 2005, 43 faculty and staff scored 382 pieces of student work (239 for communications outcome and 143 for critical thinking outcome). 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 (25%) minimizing variability. 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. Any artifacts with total scores differing by three or more after the third reader, were dropped from 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 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.

Communications Artifacts

For the 239 Communications artifacts, the scores of a little over half (52.3%) were within acceptable reliability limits (Table 1). Third readers scored 114 artifacts. When factoring in the scores of the third reader, the reliability rate increased to 93.7% (Table 2). There were 15 artifacts that were thrown out of the outcome performance analysis.

Critical Thinking Artifacts

For the 143 critical thinking artifacts, nearly 70% of the first two readers' scores were within a three-point difference (Table 3). The reliability rate this semester represents an improvement over fall 2004 of about 5 percentage points.

There were 45 (31.5%) critical thinking artifacts that needed a third reader. With the addition of the third reader, the reliability rate increased to 94.5% (Table 4). The eight scores that were outside the acceptable reliability limit will be dropped from the outcome performance analysis.

NMC's goal is to increase the reliability rate through training, reader norming sessions, and rubric improvements. 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. Ways to improve first round inter-reader reliability will be researched.

Table 1. Difference between scorer 1 and scorer 2: Communications			
Difference	Frequency	Percent	Cumulative Percent
0	32	13.4	13.4
1	51	21.3	34.7
2	42	17.6	52.3
3	38	15.9	68.2
4	30	12.6	80.8
5	19	7.9	88.7
6	14	5.9	94.6
7	3	1.3	95.8
8	5	2.1	97.9
9	3	1.3	99.2
11	1	.4	99.6
12	1	.4	100.0
Total	239	100.0	

Table 2. Difference between scorer 1 and scorer 3: Communications			
Difference	Frequency	Percent	Cumulative Percent
0	65	27.2	27.2
1	91	38.1	65.3
2	68	28.5	93.7
3	12	5.0	98.7
4	3	1.3	100.0
Total	239	100.0	

Table 3. Difference between scorer 1 and scorer 2: Critical Thinking			
Difference	Frequency	Percent	Cumulative Percent
0	34	23.8	23.8
1	38	26.6	50.4
2	26	18.2	68.6
3	17	11.9	80.5
4	13	9.0	89.5
5	4	2.8	92.3
6	4	2.8	95.1
7	4	2.8	97.9
8	1	0.7	98.6
9	1	0.7	99.3
10	1	0.7	100.0
Total	143	100.0	

Table 4. Difference between scorer 1 and scorer 3: Critical Thinking			
Difference	Frequency	Percent	Cumulative Percent
0	43	30.1	30.1
1	51	35.7	65.8
2	41	28.7	94.5
3	5	3.4	97.9

4	3	2.1	100.0
Total	143	100.0	