



Office of Institutional Research

To: Scholarship Action Group
 CC: Curriculum Committee & ESIMT
 From: Darby Hiller
 Date: August 22, 2006
 Subject: Inter-reader Reliability Study (Spring 2006)

In May, 2006, 22 faculty and staff scored 247 pieces of student work to measure progress on the Communications general education 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. 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 247 artifacts, 59.5% were within acceptable reliability limits (Table 1), an improvement over last years 52.3%. Third readers were required to score 100 artifacts.

Difference	Frequency	Percent	Valid Percent	Cumulative Percent
0	38	15.4	15.4	15.4
1	61	24.7	24.7	40.1
2	48	19.4	19.4	59.5
3	35	14.2	14.2	73.7
4	32	13.0	13.0	86.6
5	12	4.9	4.9	91.5
6	15	6.1	6.1	97.6
7	3	1.2	1.2	98.8
8	1	.4	.4	99.2
9	2	.8	.8	100.0
Total	247	100.0	100.0	

When considering the scores of the third reader, the reliability rate increased to 95.1% (Table 2). There were 12 artifacts that were outside the acceptable reliability limits and as such were thrown out of the analysis. The final outcomes analysis will include 235 artifacts.

Table 2. Difference between reader's scores when third reader considered				
Difference	Frequency	Percent	Valid Percent	Cumulative Percent
0	53	21.5	21.5	21.5
1	107	43.3	43.3	64.8
2	75	30.4	30.4	95.1
3	10	4.0	4.0	99.2
4	2	.8	.8	100.0
Total	247	100.0	100.0	

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. Table 3 shows the trends in inter-reader reliability over the semesters in which it was tracked.

Table 3. Inter-reader reliability history	2004	2005	2006
Number of artifacts scored	49	239	247
First two reads	75.5%	52.3%	59.5%
Addition of third read	100%	93.7%	95.1%
Number of artifacts thrown out of analysis	0	15	12

The data for this analysis are available in the Office of Institutional Research.