

## Writing Assessment Spring 2009

### Pre-Post Results

#### *Quality of Assessment Design*

Sixty-nine pairs of essays were rated by four raters. Each pair was written by the same student, one at the beginning of the semester (PRE) and one at the end (POST). The essays were randomly assigned to raters so that one third of the essay pairs was rated by the same two raters and one third of the pairs was rated by the other two raters. (A mixed design was achieved for the remaining third of the essay pairs, with the PRE papers rated by two raters and the POST papers rated by the other two raters.)

This assessment design allowed an assessment of the overall reliability of the scores by rater pair. In addition, comparing reliability across the two pairs of raters gives us information about the feasibility of the assessment design using two raters to score PRE and POST papers with this rubric. Reliability indexes for the two pairs of raters were 0.44 and 0.82.

The analysis revealed that, for the rater pair that achieved the higher reliability, most of the variance in scores was due to individual differences across students. For the other rater pair, there was as much variance attributable to differences between the raters as there was to differences across students. This result suggests that the design is highly dependent upon the particular raters involved. Suggestions for improving the reliability of the assessment design include:

- Continuing faculty discussion of the rubric criteria and expectations for student work.
- Review and discussion of examples of student work representing each score level for each scoring criterion (e.g., A-, B-, C-, and D-level style examples).
- Additional practice sessions in which faculty use the rubric to score the same pair of PRE-POST essays and discuss results.
- Revising the essay prompts. The analysis suggested that change over time contributed quite a bit to the variance in scores. However, another component of the variance in scores was due to the fact that this change over time was judged differently by the raters. This was true for both rater pairs. It is difficult to pinpoint the reason for such a discrepancy, but it is possible that some raters were influenced by the essay prompt (knowing which was at the beginning of the semester and which was at the end) while others were not.

For the purpose of the program assessment results, then, only the scores that were found to have acceptable reliability were analyzed.

#### *Program Assessment Results*

The performance levels on the rubric were labeled A, B, C, and D, and these have been recoded to the following conventional point values: A = 4, B = 3, C = 2, and D = 1. Average scores assigned by one rater pair for 23 pairs of PRE-POST essays are presented in the table below:

<b>Rubric Category</b>	<b>PRE</b>	<b>POST</b>	<b>Difference (POST – PRE)</b>
Content	2.97	2.98	0.01
Structure	2.63	2.96	0.33
Style	2.80	2.72	-0.09
Convention	2.59	2.67	0.09
Overall	2.93	2.91	-0.02

Two correlations were examined in order to provide validity evidence for the scores obtained in this writing assessment. PRE scores (averaged across two raters) and SAT writing scores were positively and strongly correlated ( $r = .88$ ). In addition, POST scores (averaged across two raters) and students' grades in the course were positively and moderately correlated ( $r = .47$ ). Note that the moderate correlation between POST scores and course grades is to be expected because the course grade presumably represents more than just writing ability as measured by this rubric. These results provide evidence that the scores assigned in this writing assessment are consistent with other measures of students' writing ability.

The following conclusions may be drawn from this writing assessment:

- Overall, students are writing at the B level at both the beginning and end of the semester.
- Improvement over time was observed for the rubric category Structure, with scores increasing moderately from approximately the B- to the B level.
- The Overall scores and scores in the other rubric categories did not increase.
- Content was the highest rated category (at the B level).
- Convention was the lowest rated category (at the B- level).

## Research Paper Results

### *Quality of Assessment Design*

Sixty-seven research papers were rated by three raters. Each paper was rated by two of the raters, with each rater pair rating approximately one third of the papers. This design allowed an assessment of the overall reliability of the scores by rater pair. In addition, comparing reliability across the three pairs of raters gives us information about the feasibility of the assessment design using two raters to score research papers with this rubric. Reliability indexes for the three rater pairs were 0.53, 0.75, and 0.90. These widely varying indexes suggest that the design is highly dependent upon the particular raters involved. Suggestions for improving the reliability of the assessment design include:

- Continuing faculty discussion of the rubric criteria and expectations for student work
- Review and discussion of examples of student work representing each score level for each scoring criterion (e.g., more than satisfactory, satisfactory, and less than satisfactory critical thinking examples)
- Additional practice sessions in which faculty use the rubric to score the same research paper and discuss results

For the purpose of the program assessment results, then, only the scores that were found to have acceptable reliability were analyzed.

### *Program Assessment Results*

Average scores across two rater pairs for 43 of the research papers are presented in the table below:

<b>Rubric Category</b>	<b>Average</b>
Rhetorical Knowledge	1.90
Critical Thinking	2.00
Polished Prose	1.73
Ethical Approach	1.67
Total	1.83

In order to provide validity evidence for the scores obtained in this writing assessment, correlations between research paper grades (classroom assessment) and the writing assessment scores were examined. Correlations were moderate, ranging from .46 for the critical thinking component to .57 for the overall score. Moderate correlations are to be expected since classroom grades were assigned on the basis of different rubrics and since the raters in this writing assessment were different from the classroom faculty member.

Students show strengths in rhetorical knowledge and critical thinking, as these scores are at the satisfactory level of performance. However, overall scores and scores for polished prose and ethical approach are slightly below satisfactory.