

## **GENERAL EDUCATION ASSESSMENT**

### **Critical & Analytical Thinking**

**Summer 2016**

#### **Method**

Instructors of general education courses designed to address General Education Goal #3 (Critical & Analytical Thinking) were invited to participate in the assessment project at the beginning of the Spring 2016 semester. Five instructors, representing five departments across two colleges, agreed to collect student assignments/artifacts during the Spring 2016 semester that could then be assessed using AAC&U's Value Rubric for Critical Thinking and/or Value Rubric for Inquiry Analysis). Some additional artifacts were submitted by other instructors who chose not to participate in the actual assessment process. One set of these artifacts were included in the final sample resulting in a set of six different general education classes for use in the final assessment process – Computer Science, Economics, Earth & Space Sciences, Geography, Physics, and Political Science. All courses were 100 level courses.

Training and norming sessions occurred early in the Summer of 2016. During these sessions, it became apparent that neither of the two VALUE rubrics were appropriate, by themselves, for scoring the variety of artifacts that had been collected. Thus, instructors chose the rubric (or parts of one of the rubrics) that were appropriate for use with their artifacts. They then worked with all of the other participants to help them understand how to best review and interpret their class artifacts with the chosen rubric. Three participants chose components of the Inquiry and Analysis VALUE rubric (Economics, Earth & Space Sciences, and Geography). Two participants chose components of the Critical Thinking VALUE rubric (Computer Science and Political Science). Attempts to assess artifacts from the Computer Science class proved very difficult. Despite repeated efforts to reach a minimum level of agreement, it quickly became apparent to the group that specific computer coding knowledge was needed to adequately assess student performance. Thus, it was decided that the instructor of the computer science course would be solely responsible for the assessment of those artifacts but only after completing the training sessions with the other group members using artifacts from the other four classes. A total of 35 artifacts from these four classes were randomly selected for scoring by all five group members and the group leader across three rounds. Following each round of scoring, initial scores were compared and levels of agreement determined. This was then followed by group discussion about each artifact and about the rubrics. Some minor changes were made to the rubric to clarify points of confusion and/or to better describe specific requirements for individual ratings.

The revised rubrics were then used to assess a total of 129 artifacts from each of the six classes described above, with some dimensions deleted from each rubric as not appropriate for our purposes. (The faculty group agreed that portions of the Critical Thinking VALUE rubric were most appropriate for the additional set of artifacts included in the final sample, i.e., the artifacts submitted for the Physics class). Four of the original six dimensions of the Critical

Thinking VALUE rubric were scored for at least one class: Explanation of Issues; Evidence; Student's Position; and Conclusions & Related Outcomes. Three of the original six dimensions of the Inquiry Analysis VALUE rubric were scored for at least one class: Existing Knowledge, Research, and/or Views; Analysis; and Conclusions.

Most of the 129 artifacts (108) were distributed among the five coders, with two coders assigned to each artifact, one as primary and one as secondary. Artifacts were scored by both coders only for those dimensions identified as appropriate for those specific artifacts. For the artifacts from the Computer Science class, only one group member served as a coder (as described above). Thus, final sample size varies by dimension.

Given this methodology, a total of 368 pairs of ratings were collected on a set of 108 artifacts. Agreement was assessed by measuring the number/percentage of rating pairs that differed by no more than 1 point on the associated 5-point rating scale (0-4).

### **Score Differences/Rater Agreement**

**Table 1: Score Differences**

Difference	Explanation of Issues	Evidence	Student's Position	Conclusions & Related Outcomes	Existing Knowledge	Analysis	Conclusions
0	25	23	28	21	22	25	26
1	18	18	13	20	30	21	30
2	1	3	3	3	12	18	8
3	0	0	0	0	0	0	0
Total	44	44	44	44	64	64	64
% Within 1 Rating Point	97.73%	93.18%	93.18%	93.18%	81.25%	71.88%	87.50%

Frequency of score differences within 1 rating point are highlighted. Perfect agreement between coders was achieved for 170 of the 368 score pairs (46.20%). **Agreement within one rating point was achieved for 320 of the 368 score pairs (86.96%).**

### **Final Scores**

Primary coder ratings were assigned as final ratings for each of the dimensions for the 108 artifacts when the two coder scores were within one rating point of each other. When discrepancies greater than one rating point occurred, a third coder determined the final ratings. For the computer science artifacts, the rating provided by the instructor was used as the final rating. A total of 431 ratings, across seven different dimensions of Critical and Analytical Thinking, were made.

**Table 2: Scores by Dimension**

	<b>Below Benchmark (0)</b>	<b>Benchmark (1)</b>	<b>Milestone (2)</b>	<b>Milestone (3)</b>	<b>Capstone (4)</b>	<b>Total</b>
<b>Explanation of Issues</b>	1	7	11	28	18	<b>65</b>
<b>Evidence</b>	2	1	14	26	22	<b>65</b>
<b>Student's Position</b>	0	1	15	22	6	<b>44</b>
<b>Conclusions &amp; Related Outcomes</b>	4	2	29	16	14	<b>65</b>
<b>Existing Knowledge</b>	0	8	21	30	5	<b>64</b>
<b>Analysis</b>	1	10	27	25	1	<b>64</b>
<b>Conclusions</b>	1	13	22	27	1	<b>64</b>
<b>Total</b>	<b>9 (2.09%)</b>	<b>42 (9.74%)</b>	<b>139 (32.25%)</b>	<b>174 (40.37%)</b>	<b>67 (15.55%)</b>	<b>431</b>

Nine of the ratings (2.09%) fell below benchmark level. 380 of the 431 ratings (88.17%) were at milestone level or higher. Across all dimensions, the most frequent rating assigned to any artifact was a rating of Milestone (2 or 3), with slightly more ratings at the higher milestone level than the lower milestone level overall.

### **Overall Findings/Interpretation**

Inter-rater agreement reached more than adequate levels for six of the seven dimensions assessed (with all above 80.00% and some well in the 90+% range). The agreement level for the Analysis dimension was somewhat lower (71.88%) than any of the others. It is unclear why the raters had a more difficult time agreeing on this specific dimension. But, what is clear from the table above is that the inter-rater agreement levels were lower on all of the three dimensions that were taken from the original Inquiry Analysis VALUE rubric than they were for the four dimensions taken from the Critical Thinking VALUE rubric. This may have something to do with the fact that, for the purposes of measuring inter-rater agreement, all artifacts assessed with the Critical Thinking VALUE rubric came from one single discipline (Political Science) and, for that matter, from one single class. Artifacts assessed with the Inquiry Analysis rubric came from three very different disciplines (Economics, Earth & Space Sciences, and Geography) and represented three very different types of assignments. It may well have been easier to reach higher levels of agreement with only a single type of artifact.

As mentioned above, a rating of Milestone (2 or 3) was most frequently assigned both within and across all dimensions (72.62%). For five of the seven dimensions assessed, the Milestone rating of 3 was most frequent while for the other two it was the Milestone rating of 2. It should also be noted that the frequency of Capstone ratings appears to be significantly higher in three specific dimensions (all from the Critical Thinking VALUE rubric): Explanation of Issues, Evidence, and Conclusions & Related Outcomes. As it turns out, these are the three dimensions rated by a single coder for the Computer Science course. Thus, this may be simply a consequence of a slightly biased rating process.

The preponderance of Milestone ratings (2 and 3) in this sample was not unexpected as all artifacts were obtained from students enrolled in 100 level General Education courses. It appears that our students are performing at an appropriate level in the area of Critical & Analytical Thinking.

### **Recommendations**

The assessment group suggests that the collection of artifacts for the assessment of General Education Goal #3, Critical & Analytical Thinking, be continued and that a group of faculty be recruited next year to engage in the assessment process once again. They further suggest that the assessment group be tasked with the development of a single revised rubric, one that includes appropriate dimensions of the two VALUE rubrics used in the current project, for use across a wide variety of artifacts from a wide range of disciplines.

# **GENERAL EDUCATION ASSESSMENT**

## **Ethical Decision Making**

**Summer 2016**

### **Method**

Instructors of general education courses designed to address General Education Goal #6 (Ethical Decision Making) were invited to participate in the assessment project at the beginning of the Spring 2016 semester. Three instructors, representing two departments within the College of Arts and Humanities, agreed to collect student assignments/artifacts during the Spring 2016 semester that could then be assessed using AAC&U's Value Rubric for Ethical Decision Making. One additional set of artifacts were included in the final sample resulting in a set of three different general education classes for use in the final assessment process from the departments of Philosophy and English. Courses were 100-200 level courses.

Training and norming sessions occurred early in the Summer of 2016. A total of 20 artifacts from these three classes were randomly selected for scoring by all group members. Following a round of scoring (10 artifacts), initial scores were compared and levels of agreement determined. This was then followed by group discussion about each artifact and about the rubric. Some minor changes were made to the rubric to clarify points of confusion, fit with institutional culture, and/or to better describe specific requirements for individual ratings.

The revised rubric was then used to assess a total of 128 artifacts from each of the three courses. There are five dimensions of the Ethical Decision Making VALUE rubric: Ethical Self-Awareness, Understanding, Different Ethical Perspectives/Concepts, Ethical Issue Recognition, Application of Ethical Perspectives/Concepts, and Evaluation of Different Ethical Perspectives/Concepts. Eighty-one (81) of the artifacts were assessed using all five dimensions of the rubric, 32 artifacts were assessed using only three dimensions (Understanding Different Ethical Perspectives/Concepts, Ethical Issue Recognition, Application of Ethical Perspectives/Concepts), 15 artifacts were assessed using only two dimensions (Ethical Self-Awareness and Evaluation of Different Ethical Perspectives/Concepts).

Two coders were assigned to each artifact, one as primary and one as secondary. Artifacts were scored by both coders only for those dimensions identified as appropriate for the specific artifact. Given this methodology, a total of 256 pairs of ratings were collected on a set of 128 artifacts. Agreement was assessed by measuring the number/percentage of rating pairs that differed by no more than 1 point on the associated 4-point rating scale (1-4).

### Score Differences/Rater Agreement

**Table 1: Score Differences**

Difference	Ethical Self Awareness	Understanding Diff. Ethical Perspective/Concepts	Ethical Issue Recognition	Application of Ethical Perspective/Concepts	Evaluation of Different Ethical Perspectives/Concepts
0	67	66	61	63	58
1	24	40	47	43	31
2	6	7	5	7	7
3	0	0	0	0	1
Total	97	113	113	113	97
% Within 1 Rating Point	93.81%	93.80%	95.57%	93.80%	91.75%

Frequency of score differences within 1 rating point are highlighted. Perfect agreement between coders was achieved for 315 of the 533 score pairs (59.09%). **Agreement within one rating point was achieved for 500 of the 533 score pairs (93.81%).**

### Final Scores

Primary coder ratings were assigned as final ratings for each of the dimensions for the 128 artifacts when the two coder scores were within one rating point of each other. When discrepancies greater than one rating point occurred, a third coder determined the final ratings.

**Table 2: Scores by Dimension**

	<b>Benchmark (1)</b>	<b>Milestone (2)</b>	<b>Milestone (3)</b>	<b>Capstone (4)</b>	<b>Total</b>
<b>Ethical Self Awareness</b>	28	52	12	5	<b>97</b>
<b>Understanding Different Ethical Persp./Concepts</b>	10	43	52	8	<b>113</b>
<b>Ethical Issue Recognition</b>	17	41	48	7	<b>113</b>
<b>Application of Ethical Persp./Concepts</b>	17	42	43	11	<b>113</b>
<b>Evaluation of Different Ethical Persp./Concepts</b>	18	47	21	11	<b>97</b>
<b>Total</b>	<b>90 (16.88%)</b>	<b>225 (42.21%)</b>	<b>176 (33.02%)</b>	<b>42 (7.87%)</b>	<b>533</b>

Ninety of the ratings (16.88%) were at benchmark level. 443 of the 533 ratings (83.12%) were at milestone level or higher. Across all dimensions, the most frequent rating assigned to any artifact was a rating of milestone (2 or 3), with slightly more ratings at the lower milestone level than the higher milestone level overall.

#### **Overall Findings/Interpretation**

Inter-rater agreement reached more than adequate levels for the five dimensions assessed (with all above the 90+% range). As mentioned above, a rating of Milestone (2 or 3) was most frequently assigned both within and across all dimensions (75.23%). The preponderance of Milestone ratings (2 and 3) in this sample was not unexpected as all artifacts were obtained from students enrolled in 100 and 200 level General Education courses. It appears that our students are performing at an appropriate level in the area of Ethical Decision Making.

**Recommendations**

The assessment group suggests that the collection of artifacts for the assessment of General Education Goal #6, Ethical Decision Making be continued to engage in the assessment process once again with a larger array of courses to ensure that revisions to the rubric can be applied to another set of artifacts. They further suggest it would be interesting to know the level of students the artifacts are obtained from to better understand the results (i.e. “x” number of freshman, sophomores, juniors, and seniors). While the GE courses are at the 100 or 200 level students in the courses may be junior/seniors so we may expect to see higher levels of scores for those individuals.