****

**General Education Assessment**

**2018**

**Goal #2: Employ quantitative concepts and mathematical models**

**Goal #4: Demonstrate the ability to think across and about disciplinary boundaries**

**August 2018**

**Goal #4 - Quantitative**

**Methods**

Four instructors representing three departments and four courses were invited to participate in the quantitative assessment team. The four instructors and classes were Dr. Michelle Caler, PHY 140; Dr. Afrand Agah, CSC 110; Dr. Lei Zhu, ECO 111; and Dr. Simon Condliffe, ECO 112. Dr. Megan Heckert from Geography and Planning acted as team lead but did not provide or score any artifacts. Each instructor provided 30 artifacts for assessment using the AACU value rubric for Quantitative Literacy. Prior to the first norming session, each instructor provide an answer key to their artifacts and some guidance as to which dimensions of the rubric were appropriate to which elements within the artifacts.

Three norming and assessment sessions were held in May and June 2018. Prior to the first meeting, team members used the AACU rubric to score one artifact from each class. During the meeting, each artifact was discussed in detail, with more specific attention to artifacts that were scored more than 1 point away from each other by raters. For each artifact, the professor was consulted in some detail about the artifact to reach consensus as to which dimensions were appropriate and to which parts of the artifact. Based on this understanding, team members then scored three additional artifacts from each class prior to the second meeting. The second meeting followed a similar process to the first and further refined the team’s understanding of how to apply the rubric to the chosen artifacts. Following the second meeting, all artifacts were scored. Team members met a third time to review the results, reflect on the process, and make recommendations.

Following is a narrative report of the team discussions and a summary of the assessment data.

**Summary of Discussions**

First Meeting

Meeting 1 considerably narrowed the dimensions being scored for all artifacts. Ultimately, it was decided that the CSC 110 artifact was appropriate for only 2 dimensions: interpretation and representation, while PHY 140 would be scored for 3: interpretation, representation, and calculation, ECO 111 would be scored for all dimensions with the exception of assumptions and ECO 112 would be assessed for interpretation, representation, calculation, and application/analysis.

The interdisciplinary nature of the team and artifacts proved somewhat challenging when it came to assessing answers that did not exactly match the answer keys provided. The PHY 140 artifact was the most challenging in this regard, in part because it was the least structured of the artifacts. However, all members agreed that despite the challenges, all artifacts were appropriate and would continue to be scored.

Second Meeting

The PHY 140 artifacts continued to be the most inconsistent in terms of scores – all four artifacts that had been judged to that point had at least one dimension with scores more than 1 point apart, and every dimension has had a 2 point discrepancy for at least one artifact. We further discussed what kinds of answers would receive each of the scores, with a particular focus on how to distinguish scores of 2 and 3 from each other. Given the particular challenges instructors faced in rating the artifact, the team briefly discussed whether or not to drop it from the assessment. Ultimately, the team felt that the updated scoring guidelines were worth pursuing and that we would re-visit this question at the third meeting if scores continued to have such large discrepancies. Dr. Caler agreed to create a new, more detailed answer key that would show alternate correct or partially-correct answers and all instructors agreed to re-rate all artifacts for the final assessment scores.

The ECO 112 discussion ultimately reached the conclusion that the communications dimension was not appropriate and would be dropped from the assessment. This was due to the fact that all of the answers were a single sentence and that as test responses, the answers don’t necessarily reflect attempts at clear communication due to the expert nature of the audience and limited time they had to provide their answers. The result of this decision was that none of the artifacts was assessed for the communication dimension.

At this point, the other two artifacts (CSC 110 and ECO 111) were consistently receiving scores within 1 point of each other on all dimensions and the team expressed no further areas of concern.

Third Meeting

Prior to the third meeting, all remaining artifacts were scored and summary statistics were prepared for the results. These statistics covered both inter-rater reliability and the actual assessment results.

*Inter-rater reliability*

With three raters scoring each individual artifact, the easiest metric to use to indicate inter-rater reliability was to note the range of scores that each artifact received. A range of 0 indicated that all raters gave the same score while a range of 1 would indicate that two raters agreed and one rater was off by one. Ranges of 2 and 3 are not as straightforward with regard to agreement as they could indicate two or three different scores overall, but they also do indicate where scores varied considerably. The percentage of artifacts with ranges of 0, 1, 2, or 3 was calculated for each dimension rather than for each artifact.

In reviewing the rating consistency, it was noted that for all dimensions of the rubric at least 86% of all scores were within 1 point of each other, which we feel is a good level of agreement and note would not have been likely without having used the two previous meetings to norm the rubrics.

*Student Scores*

We went on to discuss the student scores for each of the four dimensions. The scores were further broken down as average scores for each artifact and percent of artifacts scoring each rating 1-4.

In attempting to reflect on the scores as indicators of quantitative skills, the team raised the question of whether we are aiming for scores higher than 2 or higher than 3. If we look at all scores above 2, each dimension had over 90% of artifacts achieving this milestone, which would seem to indicate that the gen ed program is meeting its quantitative analysis goals. At least one member of the team questioned, however, whether 2 should actually be considered a milestone score and suggested that we should aim higher. If 3 were the more appropriate milestone to use as a goal, the percentage of artifacts scored in this range was closer to 70%, a result that was felt to be disappointing.

*The Assessment Process*

The bulk of our conversation focused on the assessment process. Based on the second meeting, communication was dropped as a dimension for all artifacts because of the difficulty of assessing correctness for non-subject matter experts. The suggestion was made that perhaps in the future a sub-set of assessors from within the discipline would be better able to assess artifacts for communication. One team member who was not able to attend the second meeting but who had participated the prior year noted that they had ultimately chosen to focus only on writing structure and not correctness at all as a way of getting around this issue. A concern was raised as to whether this would be a fair assessment of communication skills for artifacts where students were not necessarily writing with an expectation that they would be assessed on writing mechanics (and the concern that for exams in particular, where students might be pressed for time, their answers might not fully reflect their communication skills). Ultimately a suggestion was made that artifacts should be chosen where communication expectations were communicated to students.

This led to a general discussion of what makes an artifact appropriate for use. In addition to the concern about communication, the PHY 140 artifact was discussed as having been particularly difficult for non-experts to grade because it was a free response question. While this fit the general pedagogy of the physics class, the lack of structure made it difficult for non-expert raters when students did not always use the approach laid out in the answer key or write all of the components that fit the “best case” model. In some ways, this is the same suggestion made above for communication: that artifacts have structure and clarity about what is expected. The CSC 110 artifact was held up as a good example of an artifact where the expectations for the assignment were very clearly laid out and explicit in both form and content. All raters agreed this made it much easier to assess the degree to which the answers were correct in comparison to some of the others. That being said, the CS110 artifact was also scored for the fewest dimensions because the level of specificity of the assignment could only be easily matched to representation and interpretation.

We ended the discussion by talking about the rubric. Though no team member reported concerns about the wording of the rubric in terms of understanding the meanings for each score or the appropriateness of the dimensions, a few reflections were offered in retrospect about the meaningfulness of the results. In particular, one team member noted there did not seem to be even spacing between scores. Both 4 and 3 were interpreted as needing to be correct, but with 4 being better/more thorough than 3, while 2 was seen as indicating a much wider range as it would be used for some mistakes, which could be either big or small. In other words, scores of 2 were harder to interpret than others in terms of the quality of the results because the score represented a wider range of levels of correctness. At this point, most raters agreed that 1s and 4s were easiest to identify while the 2-3 distinction was much more difficult, while also noting that this is normal for assessment where extremes are generally easier to distinguish. There was not complete agreement that this was a major concern or that rubric revision was required, but two possible solutions were offered. The first was to shift the meanings slightly so that 4 would indicate correct, 3 would indicate mostly correct with minor errors, 2 would indicate somewhat correct with more major errors, and 1 would indicate completely off base. A second alternative was to shift to a 3-level rubric with scores of 3 for exceeding expectations, 2 for meeting expectations, and 1 for not meeting expectations. In both instances, the concern was not about the applicability of the rubric to the learning goal or the ability of assessor to determine appropriate scores, but about making the assessment scores easier to interpret. No consensus was reached as to a preference between these two suggestions or the necessity of making a change.

The topic of adapting the rubric topic came up as an area of concern only at the end of the third meeting. On several occasions during the first two meetings the team was asked about the rubric itself and felt that it worked – that the challenges had to do with artifacts that had not been designed to assess all of the dimensions on the rubric or the open-ended nature of the output which made it more difficult for non-experts to judge. In other words, the problem was not that they didn’t understand the rubric or that the rubric was inappropriate but that they didn’t understand the artifacts well enough to see gradations in correctness. Neither of these issues was seen as a concern that could be addressed by changing the rubric.

**Assessment Results**

Inter-rater Reliability

This table represents the range of scores assigned to each artifact. With 3 raters rather than 2 (as had been used in the past), the range was judged to be the easiest measure to use. A range of 0 indicates that all three raters assigned the same score to the artifact. A range of 1 indicates that 2 raters gave the same score and the third gave a score that was 1 point off. Ranges of 2 and 3 are more difficult to interpret, as they could indicate either all raters giving different scores or two raters in agreement and one rater substantially disagreeing. As noted in the meeting 3 summary, the team expressed satisfaction that these results indicate a fair process and a sufficiently high level of agreement between teammates.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dimension | Range 0 | Range 1 | Range 2 | Range 3 |
| Interpretation | 44.55% | 45.45% | 9.09% | 0.91% |
| Representation | 45.45% | 42.73% | 11.82% | 0.00% |
| Calculation | 51.11% | 35.56% | 13.33% | 0.00% |
| Application | 50.00% | 45.00% | 5.00% | 0.00% |

Assessment Scores

The official score assigned to each artifact was the median score assigned by the 3 raters. In cases of full agreement or a score range of 1, this was also the score that was assigned by the majority of raters. For larger ranges, the result was that artifacts would receive the middle score.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Interpretation* | **Artifact** | **Avg. Score** | **4** | **3** | **2** | **1** |
| CSC 110 | 3.65 | 14 | 5 | 1 | 0 |
| PHY 140 | 2.53 | 2 | 13 | 14 | 1 |
| ECO 112 | 2.57 | 11 | 3 | 8 | 8 |
| ECO 111 | 3.90 | 27 | 3 | 0 | 0 |
| **Overall** | **3.12** | **54** | **24** | **23** | **9** |
|  |  |  | 49.09% | 21.82% | 20.91% | 8.18% |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Representation* | **Artifact** | **Avg. Score** | **4** | **3** | **2** | **1** |
| CSC 110 | 3.30 | 10 | 6 | 4 | 0 |
| PHY 140 | 2.47 | 4 | 9 | 14 | 3 |
| ECO 112 | 2.67 | 9 | 3 | 17 | 1 |
| ECO 111 | 3.77 | 24 | 5 | 1 | 0 |
| **Overall** | **3.03** | **47** | **23** | **36** | **4** |
|  |  | 42.73% | 20.91% | 32.73% | 3.64% |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Calculation* | **Artifact** | **Avg. Score** | **4** | **3** | **2** | **1** |
| PHY 140 | 2.53 | 7 | 5 | 15 | 3 |
| ECO 112 | 3.43 | 20 | 3 | 7 | 0 |
| ECO 111 | 3.63 | 22 | 5 | 3 | 0 |
| **Overall** | **3.20** | **49** | **13** | **25** | **3** |
|  |  | 54.44% | 14.44% | 27.78% | 3.33% |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Application/Analysis* | **Artifact** | **Avg. Score** | **4** | **3** | **2** | **1** |
| ECO 112 | 2.87 | 7 | 14 | 7 | 2 |
| ECO 111 | 3.30 | 14 | 11 | 5 | 0 |
| **Overall** | **3.08** | **21** | **25** | **12** | **2** |
|  |  | 35.00% | 41.67% | 20.00% | 3.33% |

For all dimensions, the average artifact score was between 3 and 4, and indicator that overall students were providing correct answers and demonstrating quantitative literacy. That being said, the team was somewhat more concerned with the percentage of scores at each level rather than the average score. In that regard, the team again felt that these scores were generally positive, but with the caveat that this assessment was dependent on the target score. If a score of 2 is the milestone score, as in the current rubric, the result of 90% of artifacts at that score or higher is seen as a strong confirmation that WCU students are achieving quantitative literacy through the gen ed program. Some members of the team, however, felt that the program should be aiming higher – that we should instead see a score of 3 or higher as the goal, in which case there would be room for improvement.

**Recommendations**

* Choose artifacts with the rubric in mind to ensure that all dimensions of the rubric are covered in assessment.
* Choose artifacts with clear and explicit expectations in terms of both the form and content of student work to enable easier interpretation by non-expert faculty.
* For assessment of the communication dimension, choose artifacts where student expectations with regard to communication were made explicit.
* Consider revising the rubric with a focus on interpreting the results and evening out the spacing between scores.

**Goal #4 - Interdisciplinary**

**Methods**

Interdisciplinary (“I”) class instructors from four different departments in three colleges were invited to join the assessment team. Team members included Dr. Michele Bratina from Criminal Justice, Dr. Martha Donkor from History, Dr. Tim Lutz from Earth and Space Sciences and Dr. Jea Oh from Philosophy. Dr. Lisa Calvano from the Management Department served as the team leader. The courses represented were Forensic Mental Health (CRJ 336), Humans and the Environment (ESS 102), Introduction to Religious Studies (PHI 102) and Introduction to Women’s and Gender Studies (WOS 225). All classes were taught in person except Dr. Oh’s, which was 100% online.

The team met for the first time in spring semester for a training session on the assessment process and timeline. After the meeting each instructor selected an assignment from their class and posted 25-35 artifacts along with the assignment instructions on the General Education Assessment D2L page. A total of 110 artifacts were posted on D2L.

The team met three times over the summer. Prior to the first meeting, team members used the AACU Value Rubric for Integrative Learning to score the first two artifacts. During the first meeting team members decided to modify one of the rubric dimensions. Prior to the second meeting team members used the revised rubric to score five artifacts, including rescoring the first two. During the second meeting the team decided that the revised AACU rubric still did not reflect the experience of WCU students and modified the dimensions and simplified the scale. Using the new rubric, team members scored the remaining artifacts beginning at the sixth record and entered their scores in a Qualtrics database. Instructors did not score their on own artifacts. After the scoring was complete the team met a final time to discuss the results. Participants were also invited to complete a Qualtrics survey evaluating their experience.

The following report includes a summary of team discussions, assessment results, recommendations and a copy of the revised rubric.

**Summary of Team Discussions**

*First Meeting*

The purpose of this meeting was for team members to discuss their experience scoring the first two artifacts. Most team members submitted incomplete scores or no scores at all because they had difficulty using the AACU rubric. Everyone agreed that they enjoyed seeing what their colleagues were doing in class, but the wide variety of disciplines and assignment types made it challenging to interpret how the rubric related to each. Assignments included a mid-semester essay, a final exam question, a discussion board post and a photographic reflection. In response to this feedback, the agenda was modified to give each team member an opportunity to explain their assignment with respect to the rubric dimensions. Team members concluded that the “connection to experience” rubric dimension was particularly important for interdisciplinary courses and modified the description to reflect the importance of making connections outside of one’s major discipline.

*Second Meeting*

The team met a week later to review scores and discuss the experience of evaluating the first five artifacts using the revised rubric. Again, the scores were incomplete because instructors continued to have difficulty using the rubric. In particular, they were still unable to identify the specific assignment content required for students to demonstrate proficiency in each rubric dimension. However, everyone agreed that interdisciplinary competency was not simply naming concepts or ideas from other disciplines, but rather analyzing issues or solving problems using perspectives from multiple disciplines. The team also discussed whether students themselves know what it means for a class to be interdisciplinary and if students realize they are doing interdisciplinary work. A team member suggested creating a common syllabus statement about what it means for a class to be designated as interdisciplinary.

Team members also commented that the AACU rubric, even with modifications, did not reflect the experience of WCU students. In particular, they indicated that they could not score most artifacts higher than a one or a two because students seemed to struggle with making higher-order connections in their written work. They also expressed concern about whether students were simply saying the right words or regurgitating information instead of making real connections and “getting it.” However, one instructor pointed out that WCU students do make these connections in classroom discussions. In subsequent meetings, the team discussed why this was the case and concluded that it was a combination of student ability and course content, especially the tension that exists when classes serve multiple purposes in the curriculum.  The earth science class in the sample was the only one that was solely an “I” class and the syllabus was centered on a taking a “systems thinking” approach.  Thus, the course content was explicitly interdisciplinary.  The recommendation about the “I” sequence of classes was a direct result of the discussions about the ability of our students.

Another observation is that the students’ ability may also be dependent on the level of the course (100 versus 300) and also that many “I” classes have students in them ranging from first year to seniors.  One recommendation was to do the pre- and post- survey of “I” thinking the first semester students are admitted and the last semester they attend.  This would be a way of assessing “I” that was separate from “I” classes. If this is a competency we expect students to have, then team members suggested that it should not be assessed in a single class but rather across the curriculum.

The team agreed that the rubric needed to be overhauled to reflect the experience of WCU students and to be easy for faculty to use. Thus, the rubric was simplified to four common benchmarks of interdisciplinary competency that would be applicable to any WCU “I” class regardless of discipline or assignment: connection to experience; transfer of knowledge; expression of knowledge; and self-reflection and transformation. In addition, the scale was changed from holistic (i.e. descriptive) to analytic (i.e., categorical) in which artifacts are evaluated for whether they do not meet, meet or exceed expectations. Most team members were more familiar with analytic rubrics because they are used in their departmental assurance of learning processes. In addition, they felt that this type of rubric would be easier to use to evaluate “I” artifacts outside their home disciplines.

The new rubric still uses a Likert scale with ratings that roughly correspond to the AACU ratings (i.e., benchmark, etc.).  Because the application of the rubric was highly specific to each assignment per two rounds of discussions, the criteria was simplified to whether or not the students met, did not meet or exceeded the criteria.  The decision to simplify was made by the committee because so much of the discussion of scoring focused on how the specific descriptions in AACU scale did not apply to the artifacts and team members were leaving the values blank.  We tried to modify the first dimension and the same thing happened with values left blank.  Thus, the team thought collapsing and simplifying the dimensions would help.  They ultimately thought that if we were going to use a rubric to assess “I” that it should be something that worked in any class with any type of artifact.

*Third Meeting*

The original aim of this meeting was to discuss the assessment results, but the discussion immediately took a philosophical turn when a team member pointed out that interdisciplinary content is difficult to access because some “I” courses serve multiple purposes in the curriculum. For example, WOS 225 is a core class for the Women and Gender Studies major as well as an interdisciplinary, diversity and writing intensive class. PHL 108 and CRJ 335 are electives for their respective majors as well as an “I” and “W” classes. ESS101 was the only class represented in the sample that was purely interdisciplinary and not a required or elective class for a major. Given the need to cover multiple general education and major goals in many “I” classes and the mix of students taking them for different reasons, the challenge for faculty is how to keep the interdisciplinary focus front and center while balancing the need to cover discipline-specific knowledge. One team member noted that CAPC’s general education handbook provided helpful guidance on crafting an interdisciplinary syllabus.

Team members also expressed concern about whether evaluating artifacts from individual “I” classes using a rubric is the most effective way to assess interdisciplinary competency. Team members agreed that the ultimate goal of all “I” classes is changing the way students think about the world. Instructors described this phenomenon as “students learning to speak a language outside their home discipline and using it to dig deeper” and “bringing in a person with a history and past experience, shaking them up a little and sending them out into the world with a new perspective.” Thus, the goal of assessment should be measuring a change in the way students think.

Most team members use pre- and post- surveys in their own “I” classes and suggested that this may be an effective method for assessing a change in students’ thinking over time. They suggested that a committee of “I” instructors develop a common survey with benchmarks that would apply to any “I” class. However, if a rubric is used then the team agreed that it should be distributed prior to the start of the semester when faculty are developing their syllabi.

Given the importance of teaching students how to think across disciplines, team members also noted that a one three-credit “I” class was not sufficient and that perhaps the general education curriculum should include an I sequence. While general education is inherently interdisciplinary because it exposes students to content outside their major, “I” classes are where the synthesis of this information would most likely occur.

Overall, team members enjoyed and benefited from the experience of participating in general education assessment because they had an opportunity to talk to colleagues and develop a network outside their disciplines. They planned to share what they learned over the summer with their departments. In addition, they suggested that the university provide other professional development opportunities for small groups of “I” faculty (i.e., 6-8 people) to meet and discuss their courses.

**Assessment Results**

Some instructors on the team had difficulty juggling the system of reading artifacts posted on D2L and using Qualtrics to score them. Consequently, the data had to be cleaned up to remove duplicate and inaccurate records and records with missing values. The final sample size was 74 artifacts. With three instructors evaluating each artifact, there were 222 unique data points. The data was analyzed using Excel pivot tables to calculate the average score for each artifact across each rubric dimension and the total score for each course across each rubric dimension. The following table shows the average scores:



The rubric scale assigned four points for exceeding expectations, three points for meeting expectations, two points for not meeting expectations and one point for “does not apply.” The data is consistent with observations that team members made during the discussions. For example, WCU students overall struggle with making connections across disciplines in their written work. However, this may be the result of “I” classes having to balance a variety of learning outcomes if they are serving multiple purposes in the general education curriculum. The data indicate that students met expectations across all rubric dimensions in ESS 102, which is the only class that is purely interdisciplinary. All team members noted the high quality of student artifacts from this class. In addition, team members agreed that it was difficult to identify the reflective component of the artifacts from CRJ 336. The average score of 1.72 for this dimension is consistent with the observations.

**Recommendations:**

* Create a common syllabus statement about what it means for a class to be designated as interdisciplinary.
* Consider the implications for interdisciplinary content when “I” course serve multiple purposes in the curriculum.
* Assess interdisciplinary thinking by measuring the change in student thinking using a common pre- and post-survey instead of a rubric.
* If a rubric continues to be used, distribute it to all “I” faculty prior to the start of the semester when they are developing their syllabi. This idea is similar to what was suggested in the 2017 general education assessment report.
* Provide other professional development opportunities for “I” faculty to meet in small groups and share information. Again, this is similar to what was suggested in the 2017 general education assessment report.

|  |  |
| --- | --- |
| **Integrative Learning Rubric (FINAL)** |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Exceeds Expectations**  **(4)** | **Meets Expectations (3)** | **Does Not Meet Expectations**  **(2)** | **Does Not Apply**  **(1)** |
| **Connection to Experience**  Connects relevant experience and academic knowledge |  |  |  |  |
| **Transfer of Knowledge**  Draws information from other disciplines to solve problems or think about issues in own discipline or field |  |  |  |  |
| **Expression of Knowledge**  Moves beyond surface-level communication to express a deeper level of insight |  |  |  |  |
| **Self-Reflection and Transformation**  Understands how the individual connects to multiple systems in a self-reflective and transformative way. |  |  |  |  |