**General Education Learning Outcomes Assessment**

**Intercultural Knowledge and Competence & Quantitative Literacy**

**Spring 2016**

In the spring of 2016, the Office of General Education (GE) conducted an assessment of two of the General Education Learning Outcomes: Intercultural Knowledge and Competence (IKC) and Quantitative Literacy (QL). This was the first time we have assessed Intercultural Knowledge and Competence and the second time we have looked at Quantitative Literacy.

**Assessment Process**

The Office of General Education solicited artifacts from the departments of all of the General Education courses that had one of the two outcomes selected during the most recent General Education designation review. This resulted in the collection of 164 student artifacts (papers, exams, assignments) from faculty in 41 General Education courses. The levels of these courses (an approximation for difficulty) were distributed as follows: 20% at the 1000 level, 5% at the 2000 level, 45% at the 3000 level, and 30% at the 4000 level. Instructors were asked to submit four examples of student work – one low, two medium, and one high quality - that could demonstrate the achievement of the selected learning outcome.

The Office of General Education asked 18 faculty members on the General Education Curriculum Committee to participate in this review, and Assistant Vice President Ann Darling and Assistant Dean Mark St. Andre were also reviewers. The 20 reviewers were split into 10 teams of 2 individuals, and each team reviewed 8 artifacts, which totaled 80 artifacts for the whole assessment. There were more reviewers with expertise in Intercultural Knowledge and Competence, so seven of the teams looked at IKC artifacts and three teams looked at QL. All individuals completed all of their reviews.

As in previous reviews, the rubrics used to assess the learning outcomes were those developed by the Association of American Colleges and Universities (AAC&U) for the Essential Learning Outcomes. The ratings on the AAC&U rubrics (1-4) reflect levels of achievement across the whole University experience. A “1” is considered “baseline”, “2” and “3” are “milestones”, and “4” is “capstone.”

**Results**

Figures 1-4 summarize the overall results for the assessment of both learning outcomes. These figures show the total number of scores that were received for each criterion for each outcome. The “-1” category was used to indicate that reviewers did not think the criterion listed was applicable to the artifact they were reviewing. A score of “0” indicates they thought that there was no evidence in the artifact that the outcome was achieved. See Appendix A to view the full rubrics for each of the learning outcomes.

The distribution of scores in the figures show that students’ scores tend to average around a score of 2 or 3 and that the distribution of scores looks normal. The exception to this is the “-1” scores, which indicate that reviewers thought the artifact could not be used to score the criterion. One factor contributing to the normality of the scores is the fact that we asked instructors to give us one low, two medium, and one high quality artifact. Given that the artifacts in this analysis come from, on average, courses just below the 3000 level, and that most of General Education is taught at the 1000 to 4000 level, the fact that scores averaged around 2 or 3 seems appropriate.

Because this is the first year of assessing Intercultural Knowledge and Competence, it is not obvious how to interpret those scores without having some context or comparison group. That will be accomplished in the second year that this outcome is assessed. One thing that is clear is that there were quite a few artifacts that the reviewers thought could not be used to respond to the criteria, particularly the criteria of Verbal and Non-Verbal Skills, Attitudes-Openness, and Cultural Self-Awareness. This information will be used to help us give better guidance to faculty who are designing assignments and submitting artifacts.

The ratings for Quantitative Literacy cluster around the 2000 and 3000 level, which is reasonable given that most of the artifacts in this assessment came from 3000 level courses (67% of Quantitative Literacy student artifacts came from 3000 level courses, the rest were lower than 3000 level). In particular, students seemed to be interpreting, representing, and calculating data successfully (See Figure 3, where “3” was the most common rating for these criteria), and also communicating those findings (see Figure 4). There was also a sizable portion of ratings for Not Applicable (a “-1”).

Because Quantitative Literacy was assessed last year, we can compare the ratings in these two years to examine differences. The following figures show the percentage of ratings (to control for different number of artifacts examined in the different years) that fall into each of the score categories for all six of the Quantitative Literacy criteria for both the 2015 and 2016 reviews.

The first difference that is apparent between 2015 and 2016 is that the first three criteria – Interpretation, Representation, and Calculation – had higher rates of “Not Applicable” ratings in 2016 than in 2015. One possible reason for this is that the instructions given to those submitting artifacts was somewhat different in the two years.

In 2015 the University of Utah participated in the Multistate Collaborative to Advance Learning Outcomes Assessment (MSC) project through the State Higher Education Executive Officer’s Association (SHEEO) and AAC&U. The instructions to request Quantitative Literacy artifacts for those courses was standardized to meet the requirements of the MSC. Also, students whose artifacts were used in the MSC needed to have completed at least 90 credit hours toward the 122 required for their degrees. For this reason, the artifacts tended to come from higher level courses – all of them were 3000 level or higher, with 29% at the 4000 level and 12% at 5000.

It is challenging to compare 2015 to 2016 on criteria for which there were a lot of NA ratings. But, for those criteria where there was not a large difference in NA ratings – Application and Analysis, Assumptions, and Communication – the ratings were overall lower and tended to center around a rating of 2 as opposed to 3.

The assignments from the courses in the 2015 data (which include higher level courses because of the MSC study 90 credit requirement) seem more likely to require students to represent their data (see the NA columns in the Representation figure above), which is typically through charts, figures, etc. These assignments from higher level courses also seem more likely to have students doing actual calculation (see the NA’s in the Calculation figure).

What seems to be true is that students in higher level courses are being given assignments where they are more likely to have the opportunity to demonstrate a broader array of the outcomes. Further study of the kinds of assignments used in both of these years could provide us with ideas for how to request assignments that can be used to regularly assess all of the outcomes.

**Interrater Reliability (IRR)**

Figures 5 and 6 show two analyses of the interrater reliability (IRR) that was achieved by the pairs of raters in this study. The top bar is the average (mean) difference between the raters on all of their individual ratings. An average difference of around 2, for example, means that raters differed in their scores by an average of 2 points on the scale. These differences include the ratings of “-1” given to the NA scores. The other IRR score is the Spearman Rank Correlation between the two raters scores for each criterion.

The IRR scores for Intercultural Knowledge and Competence are obviously quite a bit lower, which is likely the result of this study being the first time we have assessed this outcome and thus the first time we have used the rubric. All of the mean differences in scores are greater than 1, and one of them is greater than 2.

The IRR for Quantitative Literacy, an outcome we have examined before, is noticeably higher. None of the mean differences between raters is larger than 1 point, and some correlations are as high as .83, which is considered moderately good agreement. This finding is very encouraging, because it demonstrates that our process is moving us in the direction of agreement on how to apply the rubrics.

**Discussion**

Our assessment of **Intercultural Knowledge and Competence** artifacts produced ratings that indicate students are accomplishing a level of achievement that we would expect from students in mostly 2000 level courses. At the same time, many criteria were given “NA” ratings because of the inappropriateness of the artifact for the criteria. In particular, the assignments did not seem capable of allowing students to demonstrate their Verbal and Non-Verbal Skills, Cultural Self-Awareness, and their Attitudes related to the Intercultural Knowledge and Competence learning outcome. Also, reviewers had very low agreement on their ratings, which may be the result of this being the first time we have assessed this outcome. We will initiate discussions with the General Education Curriculum Committee and faculty related to how we can improve the appropriateness of the assignments submitted or modify some of the criteria of the rubric.

The assessment of **Quantitative Literacy** artifacts produced ratings that indicate students in the sample were performing around the 2 or 3 milestone, which, again, seems appropriate for students in mostly 3000 level courses at the University. It was encouraging to see higher ratings for interpreting, representing, calculating, and communicating the results of quantitative analyses. In all of these categories, the modal (most common) rating was a 3 on the 0-4 scale. This pattern of responses in 2016 was the same as it was in 2015, which indicates that there is some consistency in our scores over time as well as between raters.

The Quantitative Literacy results can be stated with greater confidence because of the higher level of interrater reliability that was established between raters, indicating that there was decent agreement on the ratings. There were still a disappointing percentage of NA ratings on the Representation (.29), Calculation (.35), and Assumptions (.29) criteria. This fact should also be discussed in committee to determine if we can improve the appropriateness of artifacts or amend the criteria of the rubric.