General Education Assessment

Report for 2011-2012

Prepared by
Dr. Linda Siefert
General Education Assessment Director
Lea Bullard
Research Assistant for General Education Assessment
October 2012

This page purposely blank.

Acknowledgements

We would like to acknowledge the following people who provided information for this report:

Desiree Spearman from the Office of Institutional Research and Assessment, who provided demographich and preparedness data.

This page purposely blank.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

This report provides the results of the General Education Assessment efforts for academic year 2011 – 2012. This was the first year of the implementation of the University Studies curriculum, Phase 1. The UNCW Learning Goals were assessed within University Studies courses using AAC&U VALUE Rubrics and locally created rubrics, using the process recommended by the General Education Assessment Committee's March 2009 recommendations. Five Learning Goals were assessed using student work products from 17 courses within five University Studies components.

FINDINGS

FOUNDATIONAL KNOWLEDGE

In Fall 2011, 642 student work products in MAT 151 were scored on the ability to solve linear equations and exponential equations requiring logarithms. The linear equation was correctly solved by 91.7% of the students, with 3.3% receiving partial credit, and the exponential equation was correctly solved 61.5%, with 31.8% receiving partial credit. A greater percentage of students in the traditionally-sized sections correctly solved the linear equation, and a greater percentage of students taking the large lecture, technology-enhanced course correctly solved the exponential equation. Females also were more likely to solve the exponential equation.

INQUIRY

In Fall 2011, 339 student work products from CHM 101 and BIOL 105 were scored on the AAC&U VALUE Inquiry rubric. The percent of students scored at or above the proficiency level were: IN3 Design Process – 87.3%; IN4 Analysis – 72.6%; IN5 Conclusions – 77.0%; IN6 Limitations and Implications – 39.5%. Limitations and implications were not specifically asked for in either assignment, so it is noteworthy that half of the students at least began to acknowledge either limitations or implications.

SECOND LANGUAGE

In Fall 2012, 253 student work products from FRH 201 and SPN 201 were scored on locally-created writing and reading (Spanish only) rubrics. The percent of students in FRN 201 scored at or above the proficiency level were: Mechanical and Spelling – 71.0%; Grammar – 63.3%; Following Instructions – 77.8%; Content, Vocabulary & Style – 65.7%. The percent of students in SPH 201 scored at or above the proficiency level were: Content – 90.3%; Organization – 83.0%; Vocab. – appropriateness & variety: 52.1%; Vocab. – proper use: 76.7%; Grammar: 82.0%; Reading Comprehension – 82.6%.

DIVERSITY

In Spring 2012, 256 student work products from six courses in the Living in Our Diverse Nation component were scored on a locally-created diversity rubric. The percent of students scored at or above the proficiency level were: DV1 Factual Knowledge – 85.5%; DV2 Knowledge of Diverse Perspectives and Their Roots – 71.9%; DV3 Examining Diversity, History, and Culture – 61.8%; DV4 Evaluating Claims and Theories about Diversity: 82.6%. While assignments were not expected to address all dimensions of diversity, DV1 was addressed by all assignments, DV2 was addressed by 75%, and DV3 and DV4 were addressed by 50%. The assignments themselves were a major factor in how well students performed.

GLOBAL CITIZENSHIP

In Spring 2012, 155 student work products from six courses in the Living in Our Global Society component were scored on a locally-created global citizenship rubric. The percent of students scored at or above the proficiency level were: GC1 Factual Knowledge – 63.9%; GC2 Knowledge of Connections – 65.7%; GC3 Use of Diverse Cultural Frames – 66.5%; GC4 Tolerance of Differences – 76.2%; GC5 Ethical Responsibility – 64.7%. Assignments covered all dimensions well except for GC5 was covered by 50% of the assignments. GC5 is aligned to the UNCW Learning Goal but not aligned to any of the four Living in Our Global Society component student learning outcomes.

RECOMMENDATIONS

The following recommendations were adopted by the Learning Assessment Council on October 2, 2012:

- The General Education Assessment office will disaggregate the Inquiry data for dimension 6 to analyze possible differences between courses and/or sections.
- The LAC will distribute a “Did You Know” email to faculty with the results from this and other Inquiry studies and ask the faculty to share examples of what they are doing/might do regarding teaching the significance of limitations and implications to inquiry (IN6).
- The General Education Assessment office will work with the Department of Foreign Languages and Literatures to devise common rubrics for University Studies foreign language courses.
- The General Education Assessment office will provide individual results to each instructor that participated in the Diversity and Global Citizenship samples, along with scorer comments.
- A Director of University Studies position should be created and filled by July 1, 2013.

1. Background, Scope, and Methodology

Background and Scope

The University of North Carolina Wilmington Faculty Senate adopted nine UNCW Learning Goals in March 2009 (modified to eight learning goals in January 2011). The General Education Assessment process is based on the recommendations contained in the Report of the General Education Assessment Committee presented to the Provost and the Faculty Senate in March 2009. The Learning Assessment Council provides advice and feedback on the process, and recommendations based on the findings. For a complete background on the development of general education assessment at UNCW, see the *General Education Assessment Spring 2010 Report* (Siefert, 2010).

This report contains information on general education assessment activities for the academic year 2011 – 2012. In Fall 2011 and Spring 2012, the following learning goals were assessed: Foundational Knowledge, Inquiry, Second Language, Diversity, and Global Citizenship. This report outlines the methodology of and findings from five separate studies, and provides useful information on the abilities of UNCW students as measured through course-embedded assignments completed during their University Studies courses. This report also provides follow up information on the progress made on recommendations made last year, findings from a survey of actions taken, and new recommendations.

Methodology

For the purposes of this report, general education assessment activities in academic year 2011 – 2012 are divided into five areas: assessment of student learning in Foundational Knowledge, Inquiry, Second Language, Diversity, and Global Citizenship.

The following questions were examined:

- What are the overall abilities of students taking basic studies courses with regard to the UNCW Learning Goals of Foundational Knowledge, Inquiry, Second Language, Diversity, and Global Citizenship?
- What are the relative strengths and weaknesses within the subskills of those goals?
- Are there any differences in performance based on demographic and preparedness variables such as gender, race or ethnicity, transfer students vs. freshman admits, honors vs. non-honors students, total hours completed, or entrance test scores?
- What are the strengths and weaknesses of the assessment process itself?

UNCW has adopted an approach to assessing its Learning Goals that uses assignments that are a regular part of the course content. A strength of this approach is that the student work products are an authentic part of the curriculum, and hence there is a natural alignment often missing in standardized assessments. Students are motivated to perform at their best because the assignments are part of the course content and course grade. The assessment activities require little additional effort on the part of course faculty because the assignments used for the process are a regular part of the coursework. An additional strength of this method is the faculty collaboration and full participation in both the selection of the assignments and the scoring of the student work products.

The student work products collected for General Education Assessment are scored independently on a common rubric by trained scorers. The results of this scoring provide quantitative estimates of students' performance and qualitative descriptions of what each performance level looks like, which provides valuable information for the process of improvement. The normal disadvantage to this type of approach when compared to standardized tests is that results cannot be compared to other institutions. This disadvantage is mitigated in part by the use of the AAC&U VALUE rubrics for many of the Learning Goals. This concern is also addressed by the regular administration of standardized assessments, in particular, the CLA and the ETS Proficiency Profile, giving the university the opportunity to make national comparisons.

Assessment Tools

For the UNCW Learning Goals of Inquiry, the Association of American Colleges and Universities (AAC&U) Valid Assessment of Learning in Undergraduate Education (VALUE) rubric (Rhodes, 2010) was used. The VALUE rubrics, part of the AAC&U Liberal Education and America's Promise (LEAP) initiative, were developed by over 100 faculty and other university professionals. Each rubric contains the common dimensions and most broadly shared characteristics of quality for each dimension.

Locally created rubrics were used for assessing Second Language, Global Citizenship, and Diversity. The versions of each of the rubrics that were used in the study are located in the appendices of each chapter. Foundation Knowledge was scored as correct, partially correct, or incorrect.

Sample Selection

The sampling method used lays the foundation for the generalizability of the results. No one part of the University Studies curriculum, nor for that matter no one part of the university experience, is solely responsible for helping students meet UNCW Learning Goals. These skills are practiced in many courses. Each component of University Studies has its own student learning outcomes, and each of these outcomes is aligned to the Learning Goals. The University Studies Curriculum

Map in Appendix A displays this alignment. For General Education Assessment purposes, courses are selected that not only meet the learning goals, but are also among those that are taken by a large number of students, in order to represent as much as possible the work of "typical" UNCW students. Within each course, sections are divided into those taught in the classroom and completely online, taught by full-time and part-time instructors, and taught as honors or regular sections. Within each subgroup, sections are selected randomly in quantities that represent as closely as possible the overall breakdown of sections by these criteria. Within each section, all student work products are collected, and random samples of the work products are selected (sometimes consisting of all papers).

Prior to the start of the semester, the General Education Assessment staff meets with course instructors to familiarize them with the relevant rubric(s). Instructors are asked to review their course content and assignments, and to select one assignment that they feel fits some or all of the dimensions of the rubric(s) being used.

Each student enrolled in the selected course sections fills out a Student Work Product Cover Sheet, which acknowledges the use of their work for the purpose of General Education Assessment. These cover sheets are removed before scoring. The name and student ID information on the cover sheets are matched with student demographic information in university records for the purpose of analysis based on demographic and preparedness variables.

SCORING

Scorer Recruitment and Selection

Scorers are recruited from UNCW faculty and, in some cases, teaching assistants. A recruitment email is sent to chairs, sometimes to all university chairs, and sometimes to only chairs in selected departments (based on the Learning Goals and course content being assessed), asking them to forward the email to all full- and part-time faculty in their department. The desire is to include reviewers from a broad spectrum of departments. The intent is to give all faculty an opportunity to participate, to learn about the process and rubrics, and to see the learning students experience as they begin their programs. However, in some cases, the scoring is best done by discipline experts. It is also important to try to have a least one faculty member from each of the departments from which student work products were being reviewed. For the 2011-2012 studies, discipline-specific scorers were solicited for Inquiry and Second Language, whereas scorers were solicited from all departments for Diversity and Global Citizenship. Scorers were selected from those expressing an interest to make up a broad-based panel consisting of full-time and part-time faculty.

Scoring Process

Metarubrics, such as the VALUE rubrics, are constructed so that they can be used to score a variety of student artifacts across disciplines, across universities, and across preparation levels.

Their strength is also a weakness: the generality of the rubric makes it more difficult to use than a rubric that is created for one specific assignment. To address this issue, a process must be created that not only introduces the rubric to the scorers, but also makes its use more manageable.

Volunteer scorers initially attended a two to two-and-a-half hour workshop on one rubric (or two rubrics for the Second Language, Spanish scorers). During the workshop, scorers reviewed the rubric in detail and were introduced to the following assumptions adopted for applying the rubrics to basic studies work products.

Initial assumptions

1. When scoring, we are comparing each separate work product to the characteristics we want the work of UNCW graduates to demonstrate (considered to be Level 4).
2. Goals can be scored independently from each other.
3. Relative strengths and weaknesses within each goal emerge through seeking evidence for each dimension separately.
4. Common practice and the instructor's directions guide the scorer's interpretation of the rubric dimensions in relation to each assignment.
5. Additional assumptions will need to be made when each rubric is applied to individual assignments.

After reviewing the rubric and initial assumptions, the volunteers read and scored two to four student work products. Scoring was followed by a detailed discussion, so that scorers could better see the nuances of the rubric and learn what fellow scorers saw in the work products. From these discussions, assumptions began to be developed for applying the rubric to each specific assignment.

For all the Learning Goals other than Second Language, the work on common assignment-specific assumptions or guidelines was continued on the day of scoring (Second Language scorers scored their papers independently and not in a scoring session). Scorers were assigned to groups of two. Scoring of each assignment began with the pair scoring one student work product together and discussing their individual scores. Discussion clarified any implicit assumptions each scorer had used in scoring the first work product. From that discussion, each group created any assignment-specific assumptions that they would use for scoring the rest of the set of assignments. After completing a packet of work products, each scorer completed a rubric feedback form and turned in the assignment-specific assumptions used by the group. The feedback form asked for information on how well each rubric dimension fit the assignment and student work. It also asked for feedback on the quality criteria for each dimension. Scorers were also asked to complete an end-of-day survey to provide feedback on the entire process.

In order to measure the consistency of the application of the rubric, additional common work products were included in each packet for measuring interrater reliability.

2. FOUNDATIONAL KNOWLEDGE

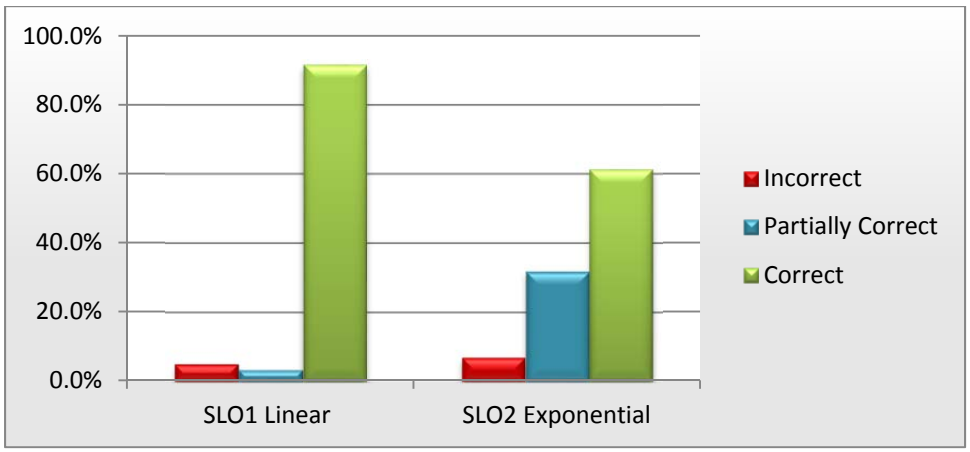
The UNCW Foundation Knowledge Learning Goal is for students to acquire foundational knowledge, theories and perspectives in a variety of disciplines. For purposes of this Learning Goal, Foundational Knowledge comprises the facts, theories, principles, methods, skills, terminology and modes of reasoning that are essential to more advanced or independent learning in an academic discipline. (UNCW Learning Goals, 2011). Eleven components of University Studies have at least one student learning outcome that is aligned to Foundational Knowledge. For this study, the course selected was from the Mathematics and Statistics component.

SUMMARY OF FINDINGS

The lead faculty teaching MAT 151 College Algebra selected the two most important student learning outcomes from the course to be assessed.

1. Students will solve linear equations.
2. Students will solve exponential equations requiring logarithms.

These student learning outcomes are aligned with the Mathematics and Statistics component SLO MS 1—employ multiple computational strategies in college-level mathematics or statistics. One test item for each of these SLOs was selected from the common course final. Section instructors applied a uniform scoring scale to the items: items were scored as correct, partially correct, or incorrect. All students taking the final were included in the sample, 642 students. Figure 2.1 provides the results.



Incorrect	5.0%	6.9%
Partially Correct	3.3%	31.8%
Correct	91.7%	61.5%

Figure 2.1 Distribution of scores for both equations

Correlation between Dimensions

The correlation between the scores on each of the two equations was .047, and it was not statistically significant. Comparing scores on the two questions, 62.0% of students who correctly solved the linear equation also correctly solved the exponential equation, similar to the 61.4% of all students who correctly solved the exponential equation. In other words, those who correctly solved the linear equation were no more likely to solve the exponential equation than those who did not.

Demographic and Preparedness Findings

The distribution of scores was statistically significantly different for males and females for exponential equations, with females answering correctly more often (64.6% vs. 56.8%, sig. = .028). To compare scores based on number of credit hours completed, two methods were used. First, students were grouped into four categories, those having completed 0 – 30 credit hours (61.1% of the sample), 31 – 60 credit hours (25.7% of the sample), 61 – 90 (9.7% of the sample), and over 90 credit hours (3.5% of the sample). Comparison of the distributions for each of the classes using the Mann-Whitney U statistic showed no statistically significant differences between them. The second method of comparison was the correlation between student scores and total hours, UNCW hours, and transfer hours. There was a statistically significant, though small, negative correlation between total hours completed and scores on the exponential equation (-.087**). SAT-Math was positively correlated with the score on the exponential equation (.197**). There were no significant correlations with GPA, ACT, or SAT-Verbal.

Comparisons between Delivery Mode

Two different modes of instruction are used to deliver MAT 151. About half of the sections are delivered in traditionally-sized classroom, and students complete and turn in practice problems, homework, and exams in paper form. The second mode of delivery combines on-line practice problems, homework, and exams with a large-lecture format (combining multiple sections) two days a week and small, TA-led practice session on Fridays. In this sample, 339 students (52.8%) were in traditionally-sized classroom sections and 303 (47.2%) were in the large-lecture, technology-enhanced sections. A greater percentage of students taking the traditionally-sized classroom-based course correctly solved the linear equation (95% vs. 88%, sig. = .000). A greater percentage of students taking the large-lecture, technology-enhanced course correctly solved the exponential equation (68% vs. 56%, sig. = .023).

Discussion

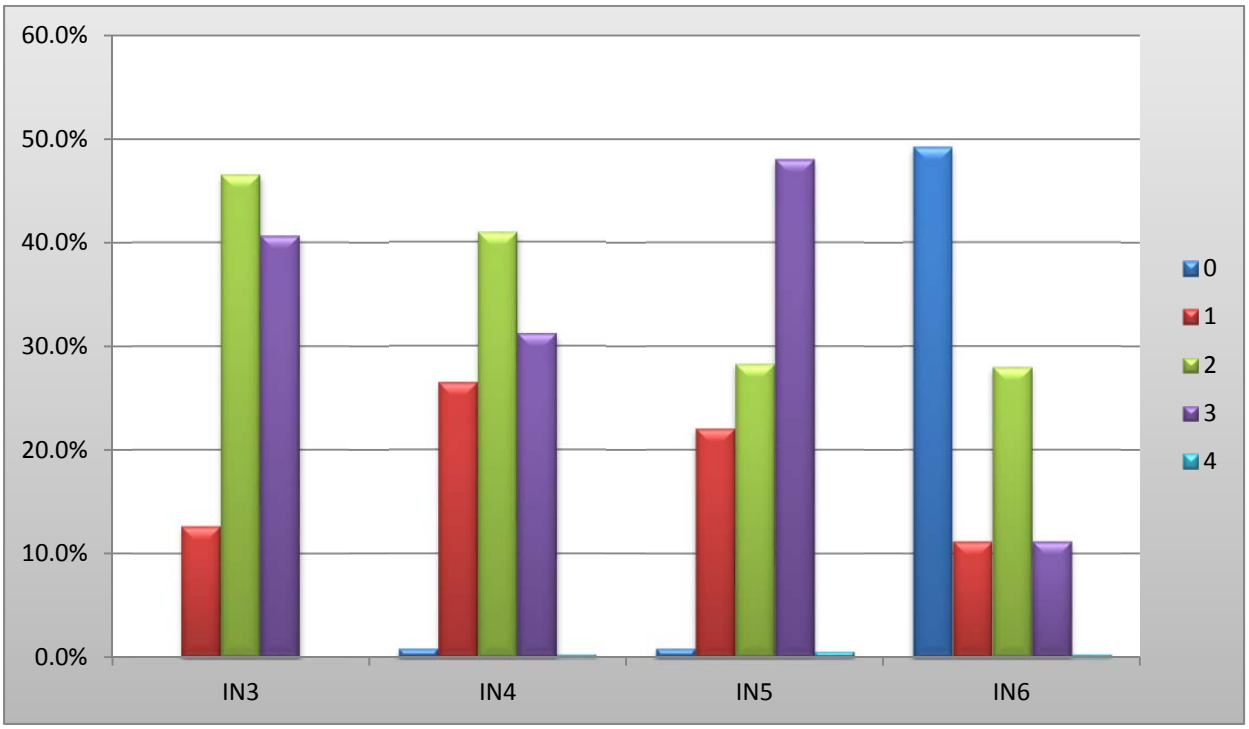
Students performed well on the linear equation, with only 5% of students' work completely incorrect. Even on the exponential equation, only 6.9% of students' work was completely incorrect. It is interesting that there were significant differences in the distribution of scores across the two delivery modes, with each group outperforming the other on one of the equations. Additionally, it would be worthwhile to compare these results to placement test results.

3. INQUIRY

The UNCW Inquiry Learning Goal is for students to engage in rigorous, open-minded and imaginative inquiry. For purposes of this Learning Goal, inquiry is the systematic and analytic investigation of an issue or problem with the goal of discovery. Inquiry involves the clear statement of the problem, issue or question to be investigated; examination of relevant existing knowledge; design of an investigation process; analysis of the complexities of the problem, clear rationale supporting conclusions; and identification of limitations of the analysis (UNCW Learning Goals, 2011). The VALUE Inquiry rubric contains six dimensions that are aligned with the UNCW description of Inquiry (see the rubric in Appendix 3.A at the end of this chapter). Twelve components of University Studies have at least one student learning outcome that is aligned to Inquiry. For this study, the courses were selected from Scientific Approaches to the Natural World.

SUMMARY OF SCORES BY DIMENSION

Twelve faculty scorers scored 339 work products from two courses from the Fall 2011 semester, CHM101 and BIOL105. The Chemistry work products were lab practicals that were completed in the classroom. The Biology work products were formal lab reports that were completed outside the class. Eighty work products (23.6%) were scored by multiple scorers. It was determined that IN1 and IN2 were not applicable for the two assignments. Figure 3.1 provides the score distributions for each dimension that was scored.



0	0.0%	0.9%	0.9%	49.3%
1	12.7%	26.5%	22.1%	11.2%
2	46.6%	41.0%	28.3%	28.0%
3	40.7%	31.3%	48.1%	11.2%
4	0.0%	0.3%	0.6%	0.3%
25th %tile	2	1	2	0
50th %tile	2	2	2	1
75th %tile	3	3	3	2
Mode	2	2	3	0
N	339	339	339	339

Figure 3.1 Distribution of Scores for Inquiry

RESULTS BY DIMENSION

IN3 Design Process

This dimension was scored for all of the assignments. Scores on this dimension were the highest of all dimensions of Inquiry (along with IN5 Conclusions). No work products failed to provide some evidence of an inquiry design (scores of 0). About one in eight work products demonstrated a misunderstanding of the methodology (scores of 1). Just under half of the work products demonstrated that critical elements of the methodology were missing, incorrectly developed, or were unfocused (scores of 2). Two in five papers showed evidence that the critical elements of the methodology were appropriately developed, though some more subtle elements were ignored or not discussed (scores of 3). No scores of 4 were assigned.

IN4 Analysis

This dimension was scored for all assignments. Scores on this dimension were in the middle range of scores for Inquiry. Fewer than one out of one hundred work products provided no evidence (scores of 0). Slightly more than one quarter of the work products listed evidence, though that evidence was not organized and/or was unrelated to the focus of the inquiry (scores of 1). Two in five work products organized evidence, but the organization was not effective in revealing important patterns, differences, or similarities (scores of 2). Almost one-third of the work products showed successful organization of evidence so that key patterns were revealed (score of 3). One work product provided evidence that was organized and synthesized to reveal insightful patterns (score of 4).

IN5 Conclusions

This dimension was deemed applicable for all assignments. The scores on this dimension were the highest, along with IN3 Design Process. Fewer than one in one hundred work products provided no conclusion (scores of 0). Just over one in five of the work products stated an ambiguous, illogical, or unsupportable conclusion (scores of 1). Over one quarter of the work products stated a general conclusion that was applicable beyond the scope of the inquiry findings (scores of 2). Almost half products stated a conclusion focused solely on the inquiry findings and that arose specifically from the inquiry findings (scores of 3 and 4).

IN6 Limitations and Implications

This dimension was viewed as applicable and was scored for all of the assignments. Scores on this dimension were the lowest of the four dimensions scored. Just under half of the work products failed to present any limitations and implications of the inquiry process (scores of 0). One out of ten work products presented limitations and implications, but those were possibly irrelevant and unsupported (scores of 1). Over one quarter of work products presented relevant and supported limitations and implications (scores of 2). One in ten work products both presented and discussed relevant and supported limitations and implications (scores of 3). One work product insightfully discussed the relevant and supported limitations and implications of the inquiry findings (score of 4).

Correlation between Dimensions

Most dimension scores were correlated with each other at the .01 or .05 level of significance, with the exception of IN3 with IN6 and IN5 with IN6. For those with correlations, the magnitudes of correlations range from .109 to .4, with the highest correlation between IN4 Analysis and IN5 Conclusions. This finding seems appropriate as a student's conclusion is developed partially through the evidence selected and discussed. See Appendix table 3.B at the end of this chapter for a complete presentation of correlation coefficients. The large and statistically significant correlations between the scores on each dimension of the rubric may

suggest a lack of independent scoring on the part of the scorers; however; they may simply represent the interdependence among all aspects of inquiry.

Demographic and Preparedness Findings

There were no statistically significant difference between the means, medians, and the score distributions of males vs. females. The samples of students with race/ethnicities other than white were too small to compare the groups. There was a statistically significant difference between the scores of transfer students vs. UNCW-start students on one dimension, IN6 Limitations and Implications. Students who began their college career at UNCW performed better at stating an inquiry conclusion drawn from the inquiry findings than did the transfer students. There were also statistically significant differences between the Honors and non-Honors students on both IN4 Analysis and IN 6 Limitations and Implications. The Honors students performed statistically higher on IN 4 Analysis and the non-Honors students performed higher on IN6.

To compare scores based on number of credit hours completed, two methods were used. First, students were grouped into four categories, those having completed 0 – 30 credit hours (61.1% of the sample), 31 – 60 credit hours (25.7% of the sample), 61 – 90 (9.7% of the sample), and over 90 credit hours (3.5% of the sample). Comparison of means (using ANOVA), medians (using Independent Samples test of medians) and distributions (using the Mann-Whitney U statistic) showed statistically significant differences between the groups for IN4, IN5, and IN6. Generally speaking, the lower classes tended to be more successful on the Inquiry dimensions than the juniors and seniors. Sophomores scored statistically higher on IN4 and IN5 than did the other classes, and freshmen scored higher on IN6. However, the sample contained only 33 juniors and 12 seniors, compared to a total of 294 freshman and sophomores. Looking at Spearman rho correlation coefficients, the number of total hours completed was positively correlated with IN5 (.246**), and transfer hours was positively correlated with IN6 (.156*).

SAT-Verbal was positively correlated with IN4 (.162**). There were no significant correlations with GPA, ACT, or SAT-Math.

Comparisons between assignment types

Work products from two types of assignments were collected for the Inquiry assessment of student learning. One was a formal lab report, which was completed outside of class. The second type of assignment was a lab practical, completed during a timed class session. There were significant statistical differences in the scores on IN3, IN5, and IN6 between the two assignment types. The in-class lab practicals scored statistically higher on IN3 Design Process and IN5 Conclusions. However, the lab reports were scored higher on IN6 Limitations and Implications. It is likely that these differences are not related to whether or not the assignments were completed in or outside of the classroom, but rather related to the emphasis given in the assignment instructions.

Interrater Reliability

For each pair of scorers, there were a number of independently-scored common papers so that interrater reliability could be assessed. Table 3.1 shows the reliability measures for Inquiry.

Table 3.1 Interrater Reliability for Inquiry

Dimension	Percent Agreement	Plus Percent Adjacent	Krippendorff's Alpha	Spearman's Rho
IN3 Design Process	61.3%	97.3%	.539	.531**
IN4 Analysis	48.0%	93.3%	.359	.358**
IN5 Conclusions	69.3%	92.0%	.617	.619**
IN6 Limitations and Implications	73.3%	93.3%	.870	.859**

Interrater reliability is a measure of the degree of agreement between scorers, and provides information about the trustworthiness of the data. It helps answer the question—would a different set of scorers at a different time arrive at the same conclusions? In practice, interrater reliability is enhanced over time through scorer discussion, as well as through improvements to the scoring rubric. Percent Agreement, Percent Agreement Plus Adjacent, and Krippendorf's Alpha measure scorer agreement. Spearman's Rho measures consistency between scorers. The UNCW benchmarks are .67 for Krippendorff's alpha and .7 for Spearman's rho. See Appendix B of the General Education Assessment 2012 Report for a more complete discussion of these statistics and the determination of benchmark levels.

Comparing the results of the reliability indices for this study to the benchmark of .67 for Krippendorff's alpha and .7 for Spearman's rho, there is one dimension of the rubric that meets these standards, IN6 Limitations and Implications. Looking at percent adjacent (that is, the scores that were within one level of each other), we find that all dimensions had greater than 90% of scores within one level of each other. Given that it was the second use of the rubric, and the first with these groups of scorers, the IRR results are very good.

Scorer Feedback on Rubric and Required Assumptions

Scorer opinion about the fit of the rubric dimensions to the assignments was mixed. For no dimension did all scorers agree that it fit equally well with the assignment; there was generally a split in scorer opinion between "fit well" and "fit with assumptions". For two dimensions, IN3 Design Process and IN5 Conclusions, scorer opinion was mixed between "fit well", "fit with assumptions", and "did not fit".

In response to the open-ended questions about the rubric dimensions that did not fit the assignment, there were two major themes of responses: those about IN3 Design Process and about IN5 Conclusions. For Design Process, scorers indicated that it proved difficult to score the students' design processes since they had been given them in their assignments. Comments included "students probably didn't realize they could add to it" and "maybe level one should be

if they just restated the methods they were given." For Conclusions, one scorer remarked that the Level 1 score "sets the bar too low" and "reads more like the description of a zero." Two other scorers mentioned that the rubric currently only describes an incorrect conclusion, and does not provide an option for a correct but not complete conclusion.

Scorers also provided feedback on what improvements could be made to the rubric. One scorer suggested that the rubric could be improved to the point where there would be no need for assumptions. Another suggested that the criteria be made more specific by defining what is meant by words such as "sufficient". Finally, two scorers proposed new dimensions for the rubric: a quantitative evidence criteria and a graphic representation dimension.

DISCUSSION

This was the third study using the finalized VALUE Inquiry and Analysis Rubric, and the first using it for lab reports in the sciences. (In Spring 2010 the rubric was used to assess work from ENG 201 and PSY 105, and in Spring 2011 it was piloted with a small sample of senior work from PLS 401.) Table 3.2 shows the percent of work products scored at a level 2 or higher and the percent of work products scored at a level 3 or higher for each dimension.

Table 3.2 Inquiry Percent of Sample Scored at Least 2 and at Least 3

Dimension	% of Work Products Scored 2 or higher	% of Work Products Scored 3 or Higher
IN1 Topic Selection	NA	NA
IN2 Existing Knowledge	NA	NA
IN3 Design Process	87.3%	40.7%
IN4 Analysis	72.6%	31.6%
IN5 Conclusions	77.0%	48.7%
IN6 Limitations and Implications	39.5%	11.5%

Scores on IN3, IN4, and IN5 were strong for 100-level courses, with 73% to 87% of the work products meeting or exceeding the benchmark level 2. The fact that scores were low on IN6 Limitations and Implications (with 49.3% of the work products scored a zero) is not a true measure of student abilities. Rather, this reflects the fact that both lab assignments asked for conclusions but did not specifically direct students to discuss limitations of the experiment or implications of the results. Therefore, it's worth noting that half of the students did at least begin to acknowledge either limitations or implications. In the Spring 2011 senior-level pilot, 71.4% of the students did not address this dimension, and faculty scorers noted that they felt that this dimension was not being given adequate attention in courses. On the other hand, in the Spring 2010 study, 54.1% of the general education students met or exceeded the level 2 benchmark. This seems to indicate that students are capable of performing well on this dimension when it is clearly asked for, and most probably, discussed in class.

The fact that IN1 and IN2 were not applicable to these lab assignments is not necessarily problematic. The pedagogy of scientific inquiry uses a guided approach, in which students are introduced to the inquiry process using well-established laboratory questions. This does point out the importance of including assignments in the majors that require students to perform all steps in the process.

The VALUE Inquiry and Analysis rubric contains the fairly common set of inquiry dimensions. Yet scorers noted in their feedback some difficulties with fitting the assignments to the rubric. In both assignments, the design process, or methodology, was given to students, so students were be assessed on their ability to recall and implement the process, which is different from creating an appropriate process. However, with that assumption, the quality criteria were applicable for scoring. The Conclusions dimension was the other dimension that led to scorer feedback. The quality criteria for this dimension are problematic, and they will be modified for the next study. The last dimension, Limitations and Implications, was not asked for directly in either assignment, and could have been removed as not applicable. However, as mentioned above, it is a positive sign that half of the students began to address either limitations or implications within their conclusions.

The Inquiry rubric and correlations table are located in the following appendices.

APPENDIX 3.A INQUIRY RUBRIC

The VALUE rubrics were developed by teams of faculty experts representing colleges and universities across the United States through a process that examined many existing campus rubrics and related documents for each learning outcome and incorporated additional feedback from faculty. The rubrics articulate fundamental criteria for each learning outcome, with performance descriptors demonstrating progressively more sophisticated levels of attainment. The rubrics are intended for institutional-level use in evaluating and discussing student learning, not for grading. The core expectations articulated in all 15 of the VALUE rubrics can and should be translated into the language of individual campuses, disciplines, and even courses. The utility of the VALUE rubrics is to position learning at all undergraduate levels within a basic framework of expectations such that evidence of learning can by shared nationally through a common dialog and understanding of student success.

Definition

Inquiry is a systematic process of exploring issues, objects or works through the collection and analysis of evidence that results in informed conclusions or judgments. Analysis is the process of breaking complex topics or issues into parts to gain a better understanding of them.

Framing Language

This rubric is designed for use in a wide variety of disciplines. Since the terminology and process of inquiry are discipline-specific, an effort has been made to use broad language which reflects multiple approaches and assignments while addressing the fundamental elements of sound inquiry and analysis (including topic selection, existing, knowledge, design, analysis, etc.) The rubric language assumes that the inquiry and analysis process carried out by the student is appropriate for the discipline required. For example, if analysis using statistical methods is appropriate for the discipline then a student would be expected to use an appropriate statistical methodology for that analysis. If a student does not use a discipline-appropriate process for any criterion, that work should receive a performance rating of "1" or "0" for that criterion.

In addition, this rubric addresses the **products** of analysis and inquiry, not the **processes** themselves. The complexity of inquiry and analysis tasks is determined in part by how much information or guidance is provided to a student and how much the student constructs. The more the student constructs, the more complex the inquiry process. For this reason, while the rubric can be used if the assignments or purposes for work are unknown, it will work most effectively when those are known. Finally, faculty are encouraged to adapt the essence and language of each rubric criterion to the disciplinary or interdisciplinary context to which it is applied.

Glossary

The definitions that follow were developed to clarify terms and concepts used in this rubric only.

- Conclusions: A synthesis of key findings drawn from research/evidence.
- Limitations: Critique of the process or evidence.
- Implications: How inquiry results apply to a larger context or the real world.

INQUIRY AND ANALYSIS VALUE RUBRIC

for more information, please contact value@aacu.org

Definition

Inquiry is the ability to know when there is a need for information, to be able to identify, locate, evaluate, and effectively and responsibly use and share that information for the problem at hand. – The National Forum on Information Literacy

Evaluators are encouraged to assign a zero to any work sample or collection of work that does not meet benchmark (cell one) level performance.

	Capstone 4	**Milestones** 3	 2	**Benchmark** 1	**Score**
Topic selection	Identifies a creative, focused, and manageable topic that addresses potentially significant yet previously less-explored aspects of the topic.	Identifies a focused and manageable/doable topic that appropriately addresses relevant aspects of the topic.	Identifies a topic that while manageable/doable, is too narrowly focused and leaves out relevant aspects of the topic.	Identifies a topic that is far too general and wide-ranging as to be manageable and doable.	
Existing Knowledge, Research, and/or Views	Synthesizes in-depth information from relevant sources representing various points of view/approaches.	Presents in-depth information from relevant sources representing various points of view/approaches.	Presents information from relevant sources representing limited points of view/approaches.	Presents information from irrelevant sources representing limited points of view/approaches.	
Design Process	All elements of the methodology or theoretical framework are skillfully developed. Appropriate methodology or theoretical frameworks may be synthesized from across disciplines or from relevant subdisciplines.	Critical elements of the methodology or theoretical framework are appropriately developed, however, more subtle elements are ignored or unaccounted for.	Critical elements of the methodology or theoretical framework are missing, incorrectly developed, or unfocused.	Inquiry design demonstrates a misunderstanding of the methodology or theoretical framework.	
Analysis	Organizes and synthesizes evidence to reveal insightful patterns, differences, or similarities related to focus.	Organizes evidence to reveal important patterns, differences, or similarities related to focus.	Organizes evidence, but the organization is not effective in revealing important patterns, differences, or similarities.	Lists evidence, but it is not organized and/or is unrelated to focus.	
Conclusions	States a conclusion that is a logical extrapolation from the inquiry findings.	States a conclusion focused solely on the inquiry findings. The conclusion arises specifically from and responds specifically to the inquiry findings.	States a general conclusion that, because it is so general, also applies beyond the scope of the inquiry findings.	States an ambiguous, illogical, or unsupportable conclusion from inquiry findings.	
Limitations and Implications	Insightfully discusses in detail relevant and supported limitations and implications.	Discusses relevant and supported limitations and implications.	Presents relevant and supported limitations and implications.	Presents limitations and implications, but they are possibly irrelevant and unsupported.	

APPENDIX 3.B CORRELATIONS BETWEEN INQUIRY DIMENSIONS

Spearman rho Rank Order Correlation Coefficients

Correlations

		TOTAL HOURS	UNCW GPA	ACT	SAT VERBAL	SAT MATH	IN 3	IN 4	IN 5	IN 6
TOTAL HOURS	Correlation Coefficient		-.221**	.073	-.027	.057	.081	.035	.246**	-.599**
	N		252	94	256	256	339	339	339	339
UNCW GPA	Correlation Coefficient	-.221**		-.070	.030	-.018	.010	.100	-.087	.128*
	N	252		59	197	197	252	252	252	252
ACT	Correlation Coefficient	.073	-.070		.540**	.359*	.144	.119	.089	-.057
	N	94	59		43	43	94	94	94	94
SAT VERBAL	Correlation Coefficient	-.027	.030	.540**		.218**	.067	.162**	.109	.104
	N	256	197	43		256	256	256	256	256
SAT MATH	Correlation Coefficient	.057	-.018	.359*	.218**		-.010	.088	.064	-.026
	N	256	197	43	256		256	256	256	256
IN 3	Correlation Coefficient	.081	.010	.144	.067	-.010		.384**	.349**	-.023
	N	339	252	94	256	256		339	339	339
IN 4	Correlation Coefficient	.035	.100	.119	.162**	.088	.384**		.400**	.109*
	N	339	252	94	256	256	339		339	339
IN 5	Correlation Coefficient	.246**	-.087	.089	.109	.064	.349**	.400**		-.009
	N	339	252	94	256	256	339	339		339
IN 6	Correlation Coefficient	-.599**	.128*	-.057	.104	-.026	-.023	.109*	-.009	
	N	339	252	94	256	256	339	339	339	

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

4. SECOND LANGUAGE

The UNCW Second Language Learning Goal is for students to demonstrate basic proficiency in speaking, listening, writing and reading in a language in addition to English (this includes American Sign Language, but not computer languages) (UNCW Learning Goals, 2011). Writing proficiency was assessed in French, and Reading and Writing proficiency were assessed in Spanish. Rubrics were developed at UNCW by each Foreign Languages and Literatures section and are included in Appendix 4.A at the end of this chapter.

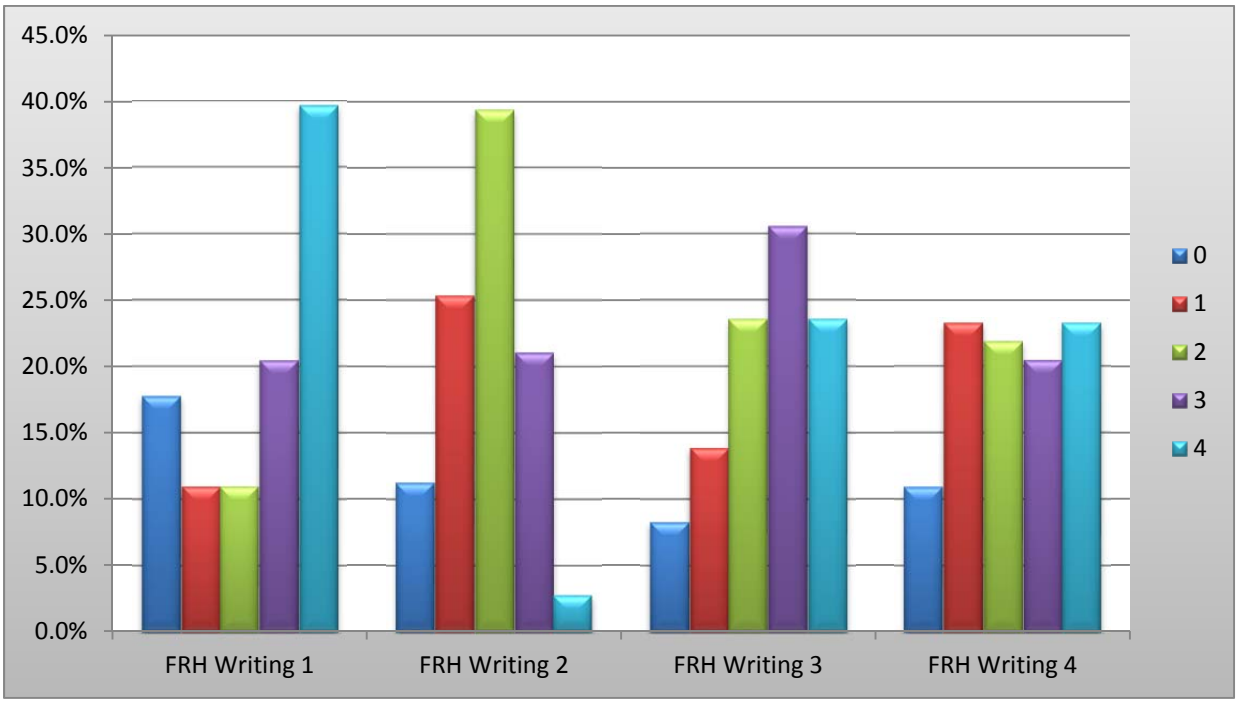
FRENCH WRITING

The rubric for writing proficiency in French consists of four dimensions, each scored on a five-point scale.

SUMMARY OF SCORES BY DIMENSION

Three faculty scorers scored 73 written work products from two courses from the Fall 2011 semester, FRH 102 and FRH 201. Fourteen work products (17.5%) were scored by multiple scorers. Figure 4.1 provides the score distributions for each dimension for work products that were scored on that dimension (work products scored as not applicable [NA] are not included).

FRENCH WRITING RESULTS BY DIMENSION FOR APPLICABLE SCORES ONLY



0	17.8%	11.3%	8.3%	11.0%
1	11.0%	25.4%	13.9%	23.3%
2	11.0%	39.4%	23.6%	21.9%
3	20.5%	21.1%	30.6%	20.5%
4	39.7%	2.8%	23.6%	23.3%
25th %tile	1	1	2	1
50th %tile	3	2	3	2
75th %tile	4	2	3	3
Mode	4	2	3	1 and 4
N	73	71	72	73

Figure 4.1 Distribution of Scores for French Writing, Applicable Scores Only

RESULTS BY DIMENSION

FRH-W1 Mechanical and Spelling

This dimension was scored for all of the assignments. Scores on this dimension were the highest of all dimensions. Less than one in five work products demonstrated many spelling and mechanical errors (scores of 0). A slightly smaller proportion, just over one in ten papers, contained several errors (scores of 1). The same ratio of papers, one in ten, received a score of two, indicating the presence of some errors. One in five papers evidenced few errors (scores of 3), and the highest number of papers, just two in five of work products showed very few errors (scores of 4).

FRH-W2 Grammar

This dimension was scored for all assignments. Scores on this dimension were in the lower range of scores. One in ten papers had grammar that was rarely correct (scores of 0). Occasionally-correct grammar was present in one of four papers (scores of 1). Two in five student work products had grammar that was usually correct (scores of 2). One in four work products contained written work with mostly correct or perfect grammar (scores of 3 and 4).

FRH-W3 Following Instructions

This dimension was deemed applicable for all assignments. The scores on this dimension were the second highest of the four dimensions. Less than one in ten students failed to follow instructions for the assignment (scores of 0). Slightly more than one in ten students did not follow all of the provided instructions (scores of 1). Almost one quarter of the written pieces demonstrated that most instructions were followed, though some questions may have been omitted or responses not fully developed (scores of 2). Over half of the student work products demonstrated that most or all of the assignment instructions had been followed (scores of 3 and 4).

FRH-W4 Content, Vocabulary & Style

This dimension was viewed as applicable and was scored for all of the assignments. Scores on this dimension were in the midrange of the four dimensions scored. Just over one in ten work products failed to communicate information or demonstrate sophistication of writing (scores of 0). Over one in five papers communicated very little information, and used simplistic and repetitive writing to do so (scores of 1). Slightly more than one in five work products demonstrated some information commutation and reliance on familiar vocabulary and sentence structure (scores of 2). One in five products communicated a lot of information and used similar and familiar sentence structure (scores of 3). Over one in five papers exhibited a varied writing style, employing a range of vocabulary and sentence structure, and communicating a breadth of information (scores of 4).

CORRELATION BETWEEN DIMENSIONS

All dimension scores were correlated with each other at the .01 level of significance. The range of correlation coefficients was .328 to .695, with FRH-W3 Following Instructions and FRH-W4 Sophistication of Writing having the highest correlation. See Appendix 4.B at the end of this chapter for a complete presentation of correlation coefficients. The large and statistically significant correlations between the scores on each dimension of the rubric may suggest a lack of independent scoring on the part of the scorers; however; they may simply represent the interdependence among all aspects of French Writing.

Demographic and Preparedness Findings

There were no statistically significant differences between the means, medians, and the score distributions of males vs. females and transfer vs. non-transfer students. The samples of students with race/ethnicities other than white were too small to compare the groups, as was the proportion of honors students to non-honors students. There were no statistically significant differences between means, medians, and score distributions between the work products from FRH 102 and FRH 201.

To compare scores based on number of credit hours completed, two methods were used. First, students were grouped into four categories, those having completed 0 – 30 credit hours (42.5% of the sample), 31 – 60 credit hours (37.0% of the sample), 61 – 90 (11.0% of the sample), and over 90 credit hours (9.5% of the sample). Comparison of means (using ANOVA), medians (using Independent Samples test of medians) and distributions (using the Mann-Whitney U statistic) showed statistically significant differences between the groups for FRH-W1 and FRH-W3. Sophomores scored higher on FRH-W1, followed by seniors, juniors, and freshmen. Freshmen scored highest on FRH-W3, followed by the senior group. Looking at Spearman rho correlation coefficients, the number of total hours completed was positively correlated with FRH-W1 (.448**).

There were no significant correlations with GPA, ACT, SAT-Verbal, or SAT-Math.

Interrater Reliability

There were a number of common student work products in each scorer's packet for scoring pairs so that interrater reliability could be assessed. Table 4.1 shows the reliability measures for French Writing.

Table 4.1 Interrater Reliability for French Writing

Dimension	Percent Agreement	Plus Percent Adjacent	Krippendorff's Alpha	Spearman's Rho
FRH-W1 Mechanical and Spelling	46.7%	60.0%	.697	.751**
FRH-W2 Grammar	40.0%	86.7%	.662	.817**
FRH-W3 Following Instructions	46.7%	73.3%	.660	.716**
FRH-W4 Sophistication of Writing	40.0%	80.0%	.502	.769**

Interrater reliability is a measure of the degree of agreement between scorers, and provides information about the trustworthiness of the data. It helps answer the question—would a different set of scorers at a different time arrive at the same conclusions? In practice, interrater reliability is enhanced over time through scorer discussion, as well as through improvements to the scoring rubric. Percent Agreement, Percent Agreement Plus Adjacent, and Krippendorf's Alpha measure scorer agreement. Spearman's Rho measures consistency between scorers. The

UNCW benchmarks are .67 for Krippendorff's alpha and .7 for Spearman's rho. See Appendix B of the General Education Assessment 2012 Report for a more complete discussion of these statistics and the determination of benchmark levels.

Comparing the results of the reliability indices for this study to the benchmark of .67 for Krippendorff's alpha and .7 for Spearman's rho, there is one dimension that meets both standards, FRH-W1 Mechanical and Spelling. All dimensions met the standard for Spearman's rho, and both FRH-W2 and FRH-W3 were close to meeting the standard for Krippendorff's alpha. Looking at percent adjacent (that is, the scores that were within one level of each other), we find that all dimensions had greater than 60% of scores within one level of each other. Interrater reliability is very good for this first use of the rubric with independent scorers.

Discussion

This was the first study using the French 102-201 Writing Assessment Rubric. Table 4.2 shows the percent of work products scored at a level 2 or higher and the percent of work products scored at a level 3 or higher for each dimension. Level 2 is the benchmark for proficiency.

Table 4.2 French Writing Percent of Sample Scored at Least 2 and at Least 3

Dimension	% of Work Products Scored 2 or higher	% of Work Products Scored 3 or Higher
FRH-W1 Mechanics and Spelling	71.0%	60.2%
FRH-W2 Grammar	63.3%	23.9%
FRh-W3 Following Instructions	77.8%	54.2%
FRn-W4 Content, Vocabulary & Style	65.7%	43.8%

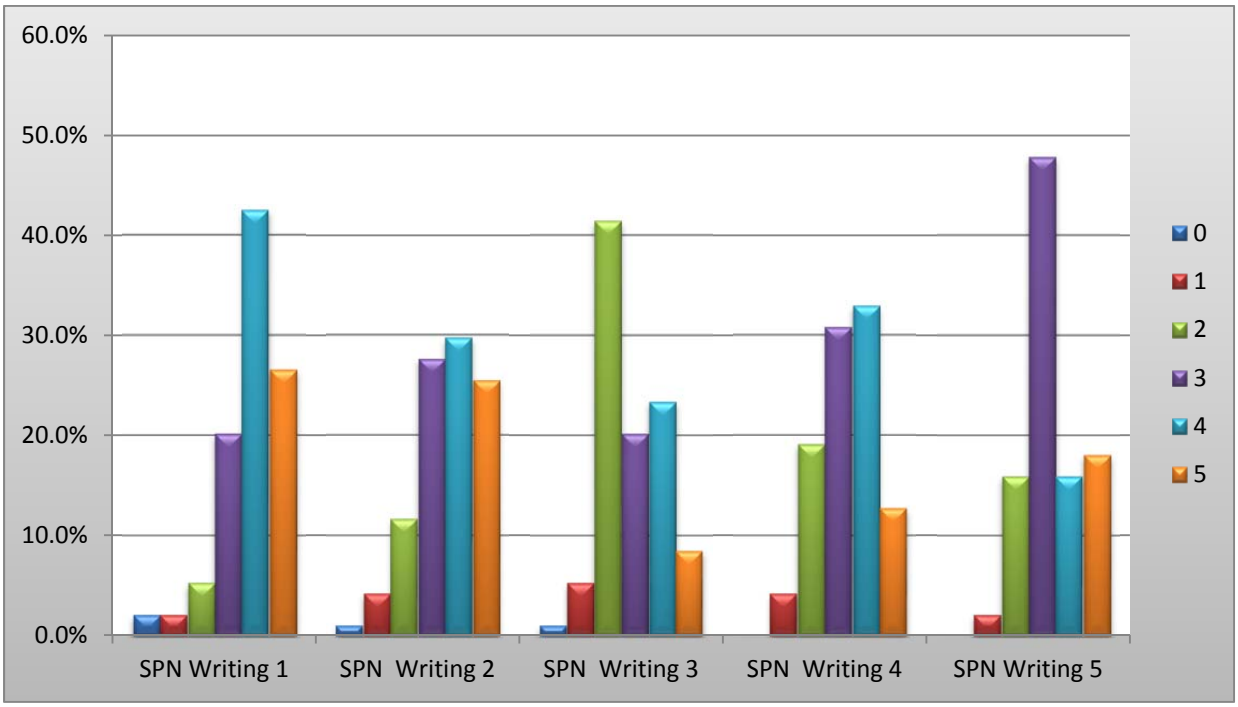
The results indicate that between two-thirds to three quarters of students attained each of the benchmark for writing proficiency by completing French 102 or 201. Interrater reliability statistics indicate that these scores are fairly reliable, and scorers reported that the process worked well and they believed it to be an appropriate way to assess students on the UNCW Learning Goals.

Spanish Writing

The rubric for writing proficiency in Spanish consists of five dimensions, each scored on a five-point scale (see Appendix 4.A at the end of this chapter).

Summary of Scores by Dimension

Seven faculty scorers scored 94 written work products from four sections of one course from the Fall 2011 semester, SPN 201. Twenty work products (21.3%) were scored by multiple scorers. Figure 4.2 provides the score distributions for each dimension for work products that were scored on that dimension (work products scored as not applicable [NA] are not included).



0	2.2%	1.1%	1.1%	0.0%	0.0%
1	2.2%	4.3%	5.3%	4.3%	2.1%
2	5.4%	11.7%	41.5%	19.1%	16.0%
3	20.4%	27.7%	20.2%	30.9%	47.9%
4	43.0%	29.8%	23.4%	33.0%	16.0%
5	26.9%	25.5%	8.5%	12.8%	18.1%
25th %tile	3	3	2	3	3
50th %tile	4	4	3	3	3
75th %tile	5	5	4	4	4
Mode	4	4	2	4	3
N	93	94	94	94	94

Figure 4.2 Distribution of Scores for Spanish Writing, Applicable Scores Only

RESULTS BY DIMENSION

SPN-W1 Content

This dimension was scored for all of the assignments, with the exception of one paper. Scores on this dimension were the highest of all dimensions. Fewer than one in twenty papers showed evidence of content being either completely unrelated to the assignment (scores of 0) or unrelated to the assignment (scores of 1). One in twenty papers had content that was too general, having little to do with the assignment (scores of 2). One in five of the written pieces contained content that, while related to the assignment, did stray from the topic at times and lacked detail (scores of 3). Slightly more than two in five papers were mostly comprehensive in terms of the

content (scores of 4), and approximately one-quarter of the papers were totally comprehensive in covering the content outlined in the assignment (scores of 5).

SPN-W2 Organization

Scores on this dimension were in the lower range of scores. About one out of twenty work products were either completely lacking in organization (scores of 0) or were poorly organized (scores of 1). Just over one in ten papers evidenced a severe lack of organization (scores of 2). Slightly more than one-quarter of the written pieces lacked some organization (scores of 3). Good organization was seen in three out of ten papers (scores of 4). One-fourth of the papers were scored as well-organized (scores of 5).

SPN-W3 Vocabulary: appropriateness and variety of lexical terms and idiomatic expressions

The scores on this dimension were the lowest of the five dimensions scored. Just over three out of fifty assignments showed either a complete lack of course-appropriate vocabulary (scores of 0) or a lack of course-appropriate vocabulary (scores of 1). Two out of five papers failed to provide sufficient use of course-appropriate vocabulary (scores of 2). One in five written pieces evidenced occasional use of appropriate lexical terms and idiomatic expressions (scores of 3), while just under one in four papers provided evidence of students' frequent use of appropriate vocabulary (scores of 4). Eight out of one hundred papers showed extensive use of lexical terms and idioms (scores of 5).

SPN-W4 Vocabulary: proper use

This dimension was viewed as applicable and was scored for all of the assignments. No papers received a zero for this dimension, which would indicate an abundance of vocabulary errors. Just over two out of fifty papers contained substantial errors (scores of 1). One in five papers evidence frequent errors in vocabulary use (scores of 2). Just under one-third of the work products showed some evidence of more-than-occasional errors (scores of 3). One-third papers contained few errors (scores of 4), and one in eight papers had almost no errors (scores of 5).

SPN-W5 Grammar

Scores for this dimension were in the mid-range of scores, and no papers received a zero score for this dimension. One in fifty papers displayed almost no evidence of correct grammar usage (scores of 1). One in six papers contained evidence of substantial grammar errors, to the extent that the intended meaning of the piece was obscured (scores of 2). Just under half of the papers contained several significant errors and/or avoidance of grammatical structures (scores of 3). One in six papers showed some grammar errors, though the meaning of the written piece was clear (scores of 4). Just under one in five papers contained only one or two significant errors (scores of 5).

Correlation between Dimensions

All dimension scores were correlated with each other at the .01 level of significance. The range of correlation coefficients was .314 to .655, with SPN-W1 Content and SPN-W2 Organization having the highest correlation. See Appendix 4.B at the end of this chapter for a complete presentation of correlation coefficients. The large and statistically significant correlations between the scores on each dimension of the rubric may suggest a lack of independent scoring on the part of the scorers; however; they may simply represent the interdependence among all aspects of Spanish Writing.

Demographic and Preparedness Findings

There were no statistically significant difference between the means, medians, and the score distributions of males vs. females and transfer vs. non-transfer students. The samples of students with race/ethnicities other than white were too small to compare the groups, as was the proportion of honors students to non-honors students.

To compare scores based on number of credit hours completed, two methods were used. First, students were grouped into four categories, those having completed 0 – 30 credit hours (46.8% of the sample), 31 – 60 credit hours (41.5% of the sample), 61 – 90 (7.4% of the sample), and over 90 credit hours (2.3% of the sample). Comparison of means (using ANOVA), medians (using Independent Samples test of medians) and distributions (using the Mann-Whitney U statistic) showed no statistically significant differences between the groups. Looking at Spearman rho correlation coefficients, the number of total hours completed was positively correlated with SPN-W3 (.235*).

There were no significant correlations with GPA and the rubric dimension scores. There was a statistically significant correlation with, ACT and SPN-W2 (.558**), and with SAT-Math and SPN-W2 (.347**), SPN-W4 (.314**), and SPN-W5 (.343**).

Comparisons between courses and instructor types

There were no significant differences in the scores between course sections taught by tenure-track faculty vs. adjunct faculty or lecturers. Likewise, there were no significant differences in the scores of the honors course section vs. the non-honors course sections.

Interrater Reliability

There were a number of common student work products in each scorer's packet by scoring pair so that interrater reliability could be assessed. Table 4.3 shows the reliability measures for Spanish Writing.

Table 4.3 Interrater Reliability for Spanish Writing

Dimension	Percent Agreement	Plus Percent Adjacent	Krippendorff's Alpha	Spearman's Rho
SPN-W1 Content	31.6%	78.9%	.128	.291
SPN-W2 Organization	42.1%	78.9%	.381	.441
SPN-W3 Vocabulary – appropriateness & variety	26.3%	57.9%	.070	.276
SPN-W4 Vocabulary – proper use	36.8%	78.9%	.057	.090
SPN-W5 Grammar	36.8%	84.2%	.424	.458*

Interrater reliability is a measure of the degree of agreement between scorers, and provides information about the trustworthiness of the data. It helps answer the question—would a different set of scorers at a different time arrive at the same conclusions? In practice, interrater reliability is enhanced over time through scorer discussion, as well as through improvements to the scoring rubric. Percent Agreement, Percent Agreement Plus Adjacent, and Krippendorf's Alpha measure scorer agreement. Spearman's Rho measures consistency between scorers. The UNCW benchmarks are .67 for Krippendorff's alpha and .7 for Spearman's rho. See Appendix B of the General Education Assessment Report 2012 for a more complete discussion of these statistics and the determination of benchmark levels.

Comparing the results of the reliability indices for this study to the benchmark of .67 for Krippendorff's alpha and .7 for Spearman's rho, there are no dimensions that meet these standards. Looking at percent adjacent (that is, the scores that were within one level of each other), we find that all but one dimension, SPN-W3 Vocabulary: appropriateness and variety of lexical terms and idiomatic expressions, had greater than 70% of scores within one level of each other.

Overall, these various measures of reliability illustrate randomness in agreement and indicate that additional norming activities are required.

Scorer Feedback on Rubric and Required Assumptions

Scorers were not as satisfied with the match between the rubric and assignments. One scorer felt that the rubrics would need to be edited for future use, while another felt that some of the assignment content needed revision to better match the rubric dimensions.

Discussion

The Spanish section has used this rubric for approximately four years, making some revisions along the way. This was the first study in which student work was scored by independent scorers. Table 4.4 shows the percent of work products scored at a level 3 or higher and the percent of work products scored at a level 4 or higher for each dimension. Level 3 is the benchmark for writing proficiency.

Table 4.4 Spanish Writing Percent of Sample Scored at Least 2 and at Least 3

Dimension	% of Work Products Scored 3 or higher	% of Work Products Scored 4 or Higher
SPN-W1 Content	90.3%	69.9%
SPN-W2 Organization	83.0%	55.3%
SPN-W3 Vocabulary – appropriateness & variety	52.1%	31.9%
SPN-W4 Vocabulary – proper use	76.7%	45.8%
SPN-W5 Grammar	82.0%	34.1%

At least 75% of the student work met the proficiency for each dimension of writing except SPN-W3 Vocabulary-appropriateness and variety, on which about half of the student work demonstrated proficiency. The two vocabulary dimensions had the lowest interrater reliability statistics. These two findings indicate that vocabulary is an area for faculty discussion. Low interrater reliability on all dimensions, coupled with scorer comments expressing dissatisfaction with the match between the assignment and the rubric and the need for additional rubric modification, also indicate areas for additional modifications.

SPANISH READING

The assessment for reading proficiency in Spanish consisted of a short reading passage that followed by five short-answer questions. Each student response was scored as either incorrect (score of 0), partially correct (score of 1), or correct (score of 2). During the initial scoring workshop, the scorers discussed and agreed on the information points required for a correct answer and guidelines for partial answers. Since all five questions measured the same dimension of reading comprehension, main idea and supporting details, scores from all five questions were added for a final score for each student work product.

SUMMARY OF SCORES BY DIMENSION

Six faculty scorers scored 86 work products from four sections of one course from the Fall 2011 semester, SPN 201. Twenty-four work products (27.9%) were scored by multiple scorers. Figure 4.3 provides the summed scores of the reading assessment.

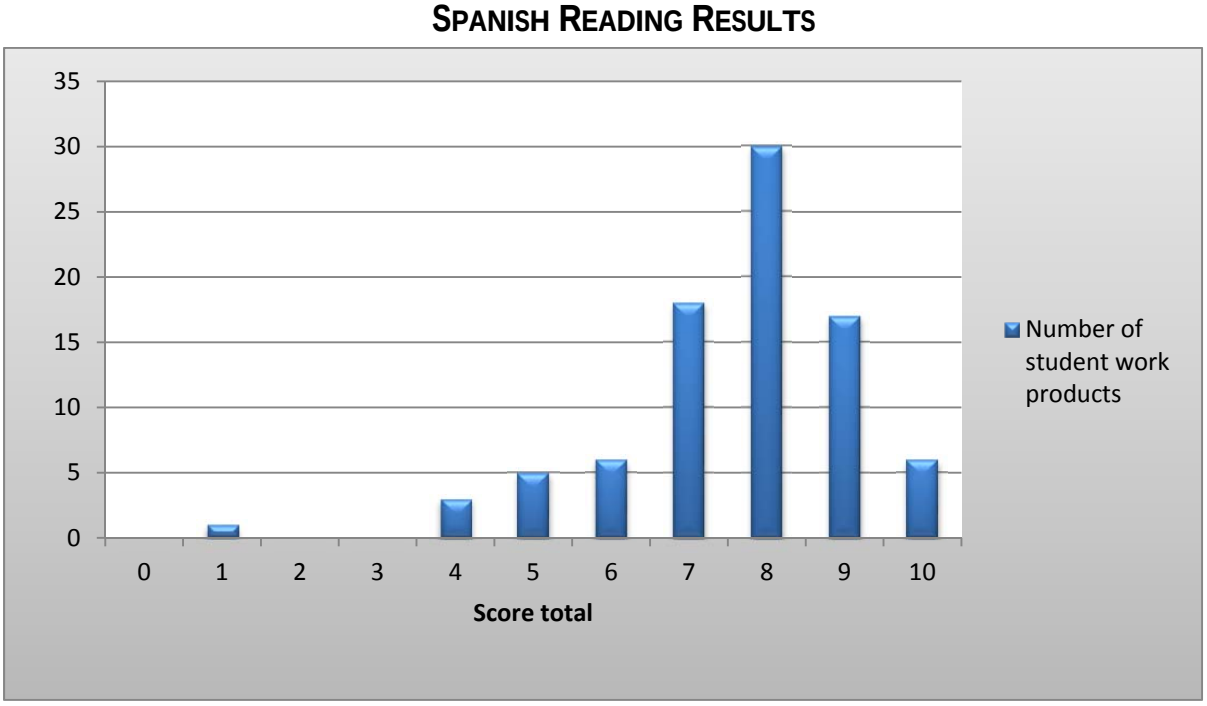


Figure 4.3 Summed Scores for Five Spanish Reading Questions

RESULTS

A total score of 70% (7 points out of the possible 10) was determined as the benchmark for reading proficiency. The majority of students, 82.6%, scored 70% or more on the Spanish reading assessment, and 61.6% scored 80% or higher. Examining individual test items, most of the short-answer responses, just fewer than 60%, were scored as correct (score of 2). About one-third of the responses to the individual questions were scored as partially correct (score of 1). Fewer than 19% of the responses were scored as completely incorrect (score of 0).

INTERRATER RELIABILITY

Table 4.5 Interrater Reliability for Spanish Reading

	Percent Agreement	Plus Percent Adjacent	Krippendorff's Alpha	Spearman's Rho
All questions	82.4%	98.4%	.765	.772**

Interrater reliability is a measure of the degree of agreement between scorers, and provides information about the trustworthiness of the data. It helps answer the question—would a different set of scorers at a different time arrive at the same conclusions? In practice, interrater reliability is enhanced over time through scorer discussion, as well as through improvements to the scoring rubric. Percent Agreement, Percent Agreement Plus Adjacent, and Krippendorf's Alpha measure scorer agreement. Spearman's Rho measures consistency between scorers. The UNCW benchmarks are .67 for Krippendorff's alpha and .7 for Spearman's rho. See Appendix B

of the General Education Assessment Report for 2012 for a more complete discussion of these statistics and the determination of benchmark levels.

Comparing the results of the reliability indices for this study to the benchmark of .67 for Krippendorff's alpha and .7 for Spearman's rho, the results meet that standard. Looking at percent adjacent (that is, the scores that were within one level of each other), we find that the scorers' marks were either in agreement or within one score level of one another 98.4% of the time.

Discussion

Over 80% of students met the proficiency score of 70%. Interrater reliability was good. The process worked well, with much discussion in the initial workshop, and the generation of consensus around the content required for correct and partially correct answers. However, two-thirds of the scorers indicated that they did not feel the process was an appropriate way to assess the learning goals, and some were not satisfied with the reading passages. In addition, it was noted that the test questions assessed only main idea and supporting details. Based on these findings, additional discussion and recommendations are warranted.

The Second Language rubrics and correlations tables are located in the following appendices.

APPENDIX 4.A SECOND LANGUAGE RUBRICS

4.A.1 FRENCH WRITING ASSESSMENT RUBRIC

Grammar and Mechanics / Proofreading

Excellent	Good	Acceptable	Poor	Unacceptable
You proofread well missing very few (0-2) spelling and mechanical errors.	You proofread well missing a few (4-5) spelling and mechanical errors.	You proofread missing some (6-8) spelling and mechanical errors.	You did not proofread well missing several (9-10) spelling and mechanical errors.	You did not proofread well missing many (>10) spelling and mechanical errors.
Your grammar is completely correct. A Native speaker would have no difficulty understanding you.	Your grammar is mostly correct, but a Native speaker would have no difficulty understanding you.	Your grammar is usually correct. A Native speaker might have occasional difficulty understanding you.	Your grammar is only occasionally correct. A Native speaker would find your writing difficult to understand.	Your grammar is rarely correct. A Native speaker would find your writing mostly incomprehensible.

Content, Vocabulary, and Style

Excellent	Good	Acceptable	Poor	Unacceptable
You follow all of the instructions: you have all the required length and address all of the questions.	You follow the instructions for the most part: you have all the required length and address most of the questions.	You follow the instructions for the most part: you have omitted some questions and your responses are not fully developed.	You do not follow all of the instructions: you have answered few of the questions and have not provided developed answers.	You fail to follow the instructions: most questions are not answered and/or required length has not been met.
You communicate a lot of information about the topics. Your writing is also interesting, because you vary and correctly use a range of vocabulary and sentence structures.	You communicate a lot of information about the topics. You correctly use most of the vocabulary and sentence structures, but there isn't much variety: you tend to repeat familiar or similar items.	You communicate some information about the topics. You rely on familiar vocabulary and items and do not vary your sentence structure.	You communicate very little information about the topics. Your writing is highly simplistic and repetitive, using the most basic and repetitive vocabulary and sentence structure.	You communicate virtually no information about the topics.

4.A.2 SPANISH WRITING ASSESSMENT RUBRIC

Content and Organization (logical sequencing, appropriate length, and comprehensiveness of information)		**Vocabulary** (appropriateness & variety of lexical items, idiomatic expressions, & use of Spanish)		**Grammar** (control of course-appropriate structures, forms, & syntax; spelling)	
5	well organized; content totally comprehensive	5	extensive use of course-appropriate vocabulary, with almost no errors	15	one or two significant errors
4	good organization; content mostly comprehensive	4	frequent use of course-appropriate vocabulary, with few errors	12	some significant errors, but meaning clear
3	some lack of organization; content strays from assignment and lacks detail	3	occasional use of course-appropriate vocabulary, with more than occasional errors	9	several significant errors and or avoidance of structures
2	severe lack of organization; content has little to do with assignment, is too general	2	insufficient use of course-appropriate vocabulary, with frequent errors	6	substantial errors; meaning is obscured
1	poorly organized; content unrelated to assignment	1	lack of course-appropriate vocabulary, with substantial errors	3	correct usage of grammar almost non-existent
0	complete lack of organization; content completely unrelated to assignment	0	complete lack of course-appropriate vocabulary, with an abundance of errors	0	correct usage of grammar non-existent

4.A.3 SPANISH READING ASSESSMENT RUBRIC

Literal Comprehension How well can you identify		**Literal Comprehension** How well can you		**Interpretive Comprehension** How well can you infer the meaning of cognates and word		**Interpretive Comprehension** How well can you infer		**Interpretive Comprehension** How well can you infer	
4	You identify all of the main ideas presented in the text.	4	You understand all of the supporting details of the text.	4	You infer the meaning of all cognates and word families.	4	You infer the meaning of many new words from context.	4	You infer the author's intent.
3	You identify most of the main ideas presented in the text.	3	You understand most of the supporting details of the text.	3	You infer the meaning of most cognates and word families.	3	You infer the meaning of most of the new words from context.		
2	You identify some of the main ideas presented in the text.	2	You understand some of the supporting details of the text.	2	You infer the meaning of some cognates and word families.	2	You infer the meaning of some new words from context.		
1	You identify few main ideas of the text.	1	You understand few supporting details of the text.	1	You infer the meaning of few cognates and word families.	1	You infer the meaning of very few words from context.		
0	You cannot identify the main ideas of the text.	0	You understand none of the supporting details of the text.	0	You cannot infer the meaning cognates and word families.	0	You do not derive the meaning of new words from context.	0	You do not infer the author's intent.
Score:		**Score:**		**Score:**		**Score:**		**Score:**	
Comments:		**Comments:**		**Comments:**		**Comments:**		**Comments:**	

APPENDIX 4.B CORRELATIONS BETWEEN SECOND LANGUAGE DIMENSIONS

TABLE 4.B.1 . SPEARMAN RHO RANK ORDER CORRELATION COEFFICIENTS FOR FRENCH WRITING

Correlations

		UNCW GPA	TOTAL HOURS	ACT	SAT VERBAL	SAT MATH	FR Writing 1	FR Writing 2	FR Writing 3	FR Writing 4
UNCW GPA	Correlation Coefficient		.393**	.399	-.280	-.052	.432**	.134	.278*	.139
	N		57	12	42	42	57	55	56	57
TOTAL HOURS	Correlation Coefficient	.393**		.088	-.172	-.084	.448**	.107	.186	.013
	N	57		15	55	55	73	71	72	73
ACT	Correlation Coefficient	.399	.088		.184	.276	.125	.059	-.021	-.116
	N	12	15		9	9	15	14	15	15
SAT VERBAL	Correlation Coefficient	-.280	-.172	.184		.403**	-.202	.012	-.054	.014
	N	42	55	9		55	55	54	55	55
SAT MATH	Correlation Coefficient	-.052	-.084	.276	.403**		-.032	-.077	.020	.028
	N	42	55	9	55		55	54	55	55
FR Writing 1	Correlation Coefficient	.432**	.448**	.125	-.202	-.032		.520**	.328**	.387**
	N	57	73	15	55	55		71	72	73
FR Writing 2	Correlation Coefficient	.134	.107	.059	.012	-.077	.520**		.497**	.671**
	N	55	71	14	54	54	71		70	71
FR Writing 3	Correlation Coefficient	.278*	.186	-.021	-.054	.020	.328**	.497**		.695**
	N	56	72	15	55	55	72	70		72
FR Writing 4	Correlation Coefficient	.139	.013	-.116	.014	.028	.387**	.671**	.695**	
	N	57	73	15	55	55	73	71	72	

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

TABLE 4.B.2. SPEARMAN RHO RANK ORDER CORRELATION COEFFICIENTS FOR SPANISH WRITING

			UNCW GPA	TOTAL HOURS	ACT	SAT VERBAL	SAT MATH	SPN Writing 1	SPN Writing 2	SPN Writing 3	SPN Writing 4	SPN Writing 5
Spearman's rho	UNCW GPA	Correlation Coefficient		.448**	-.267	-.151	.102	-.152	-.094	-.032	.105	-.034
		N		179	48	155	155	77	78	78	78	78
	TOTAL HOURS	Correlation Coefficient	.448**		-.231	-.156*	-.094	.036	.022	.235*	.052	.023
		N	179		60	177	177	93	94	94	94	94
	ACT	Correlation Coefficient	-.267	-.231		.671**	.656**	.332	.558**	-.032	.056	.133
		N	48	60		46	46	27	27	27	27	27
	SAT VERBAL	Correlation Coefficient	-.151	-.156*	.671**		.071	.110	.144	.002	.098	.031
		N	155	177	46		177	79	80	80	80	80
	SAT MATH	Correlation Coefficient	.102	-.094	.656**	.071		.199	.347**	.135	.314**	.343**
		N	155	177	46	177		79	80	80	80	80
	SPN Writing 1	Correlation Coefficient	-.152	.036	.332	.110	.199		.655**	.466**	.554**	.536**
		N	77	93	27	79	79		93	93	93	93
	SPN Writing 2	Correlation Coefficient	-.094	.022	.558**	.144	.347**	.655**		.525**	.611**	.489**
		N	78	94	27	80	80	93		94	94	94
	SPN Writing 3	Correlation Coefficient	-.032	.235*	-.032	.002	.135	.466**	.525**		.314**	.358**
		N	78	94	27	80	80	93	94		94	94
	SPN Writing 4	Correlation Coefficient	.105	.052	.056	.098	.314**	.554**	.611**	.314**		.601**
		N	78	94	27	80	80	93	94	94		94
	SPN Writing 5	Correlation Coefficient	-.034	.023	.133	.031	.343**	.536**	.489**	.358**	.601**	
		N	78	94	27	80	80	93	94	94	94	

**. Correlation is significant at the 0.01 level (2-tailed).

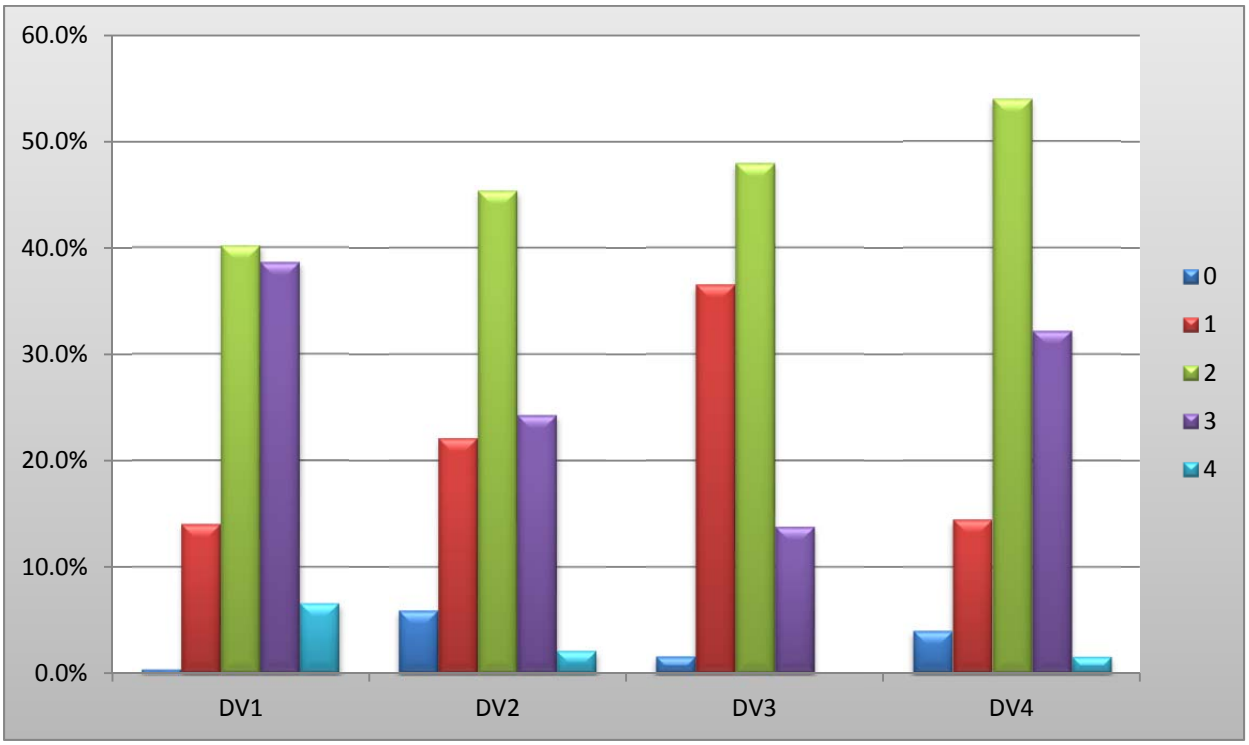
*. Correlation is significant at the 0.05 level (2-tailed).

5. DIVERSITY

The UNCW Diversity Learning Goal is for students to describe and examine the importance and implications of human diversity. For purposes of this Learning Goal, diversity constitutes the knowledge, skills and attitudes necessary to examine the importance and implications of cultural and ethnic human differences. Diversity examines the significance of historical, political, social, racial, ethnic and cultural realities through critical thinking to understand and explain their implications in human endeavors (UNCW Learning Goals, 2011). The Diversity rubric was developed in-house at UNCW and details four dimensions of knowledge related to human diversity (see Appendix 5.A at the end of this chapter for the Diversity rubric). Five components of University Studies have at least one student learning outcome aligned to Diversity. In this study, student work was sampled from the Living in Our Diverse Nation component.

SUMMARY OF SCORES BY DIMENSION

Eight faculty scorers scored 256 work products from six courses in the Spring 2012 semester: GRN 101, HST 204, MUS 117, PSY 270, PSY 271, and SOC 325. A variety of assignment types were sampled; both in-class and out-of-class work, short and long written pieces, online discussion postings and journal entries, and exam essay questions. Fifty-six work products (21.9%) were scored by multiple scorers. Not all dimensions of the rubric were applicable for a particular assignment. Only one assignment was scored on all four dimensions. However, for two courses, student work from two assignments was provided, and all four dimensions were covered across the two assignments. DV1 Factual Knowledge was aligned to all assignments. Figure 5.1 provides the score distributions for each dimension for work products that were scored on that dimension (i.e., work products for which the dimension was not applicable are not included in the totals).



0	0.4%	5.9%	1.6%	4.0%
1	14.1%	22.2%	36.6%	14.5%
2	40.2%	45.4%	48.0%	54.0%
3	38.7%	24.3%	13.8%	32.3%
4	6.6%	2.2%	0.0%	1.6%
25th %tile	2	1	1	2
50th %tile	2	2	2	2
75th %tile	3	3	2	3
Mode	2	2	2	2
N	256	185	123	132

Figure 5.1 Distribution of Scores for Diversity, Applicable Scores Only

RESULTS BY DIMENSION

DV1 Factual Knowledge

This dimension was scored for all of the assignments. Scores on this dimension were the highest along with DV4 Evaluating Theories. Less than one in two hundred work products failed to use any terminology surrounding diversity (scores of 0). One in seven work products used some terminology surrounding diversity but identified few, in any of the basic elements of a diversity issue (scores of 1). Two in five work pieces identified some of the basic elements of an issue or diversity theme, with an incomplete or inaccurate description (scores of 2). Slightly less than two in five work products accurately explained the major elements of the diversity issue of

theme at hand (scores of 3). Around seven in one hundred samples provided a comprehensive, detailed, and accurate discussion of a diversity issue (scores of 4).

DV2 Knowledge of Diverse Perspectives and Their Roots

This dimension was scored for four assignments. Scores on this dimension were in the middle range of scores for Diversity. Less than one in ten work products failed to demonstrate any evidence of knowledge of diverse perspectives and their roots (scores of 0). Almost one-quarter of the sampled products identified some elements of the perspectives of a specific social group, but did not demonstrate an awareness of societal or cultural influences on these perspectives (scores of 1). Just under half of the student work both identified some elements of the perspectives of a particular social group and provided some explanation of how culture and society influenced those perspectives (scores of 2). One quarter of work products explained the important aspects of the perspectives of the social group, and discussed in more detail how those perspectives are influenced (and continue to be influenced) by society (scores of 3). Two out of one hundred papers showed evidence of a detailed discussion about groups' perspectives and a comprehensive examination of how culture and society influence those perspectives (scores of 4).

DV3 Examining Diversity, History, and Culture

This dimension was deemed applicable for three assignments. The scores on this dimension were in the lower range of scores. One in one hundred work products failed to demonstrate any knowledge of diversity, history, and culture (scores of 0). Slightly more than one-third of the sampled papers presented a narrow set of evidence that was taken as factual without questioning in order to describe the influence of human diversity on the history and/or culture of the United States (scores of 1). Just less than one-half of the work products presented evidence that was taken mainly as factual with some, but little, questioning to support the discussion of the influence of human diversity on history and culture (scores of 2). Just over one in ten papers supplied substantial evidence that is both relevant and supportive of the influence of human diversity (scores of 3). No papers were scored a four for this dimension.

DV4 Evaluating Claims and Theories about Diversity

This dimension was viewed as applicable and was scored for four assignments. Scores on this dimension were the second highest of the dimensions scored. Less than one in twenty work products scored a zero on this dimension, indicating no evaluation of claims and theories about diversity. One in seven papers attempted to provide evidence that backs up or disputes a claim, argument, or theory regarding the interplay between diversity, identity, and experience, although with unrelated or inaccurate evidence (scores of 1). Over one-half of the work products provided some evidence for or against a claim or argument, although the argument was incomplete (scores of 2). One-third of the sampled papers presented an evidence-based, accurate, and substantially

complete argument (scores of 3) or a well-thought-out argument that acknowledged competing viewpoints (scores of 4).

Correlation between Dimensions

All dimension scores were correlated with each other at the .01 of significance, with the highest correlation of .740 between DV2 Knowledge of Diverse Perspectives and DV4 Evaluating Claims and Theories. See Appendix table 5.B at the end of this chapter for a complete presentation of correlation coefficients. The large and statistically significant correlations between the scores on each dimension of the rubric may suggest a lack of independent scoring on the part of the scorers; however; they may simply represent the interdependence among all aspects of diversity.

Demographic and Preparedness Findings

There was a statistically significant difference between the scores of transfer students vs. UNCW-start students on one dimension, DV2, Knowledge of Diverse Perspectives, with UNCW-state students scoring higher on that dimension. There were also statistically significant differences between the male and female groups on DV1 Factual Knowledge, DV2 Knowledge of Diverse Perspectives, and DV3 Examining Diversity, History, and Culture, with females scoring higher. There were no statistically significant difference between the means, medians, and the score distributions of honors vs. non-honors students. The samples of students with race/ethnicities other than white were too small to compare the groups.

To compare scores based on number of credit hours completed, two methods were used. First, students were grouped into four categories, those having completed 0 – 30 credit hours (22.7% of the sample), 31 – 60 credit hours (25.8% of the sample), 61 – 90 (34.0% of the sample), and over 90 credit hours (17.6% of the sample). Comparison of means (using ANOVA), medians (using Independent Samples test of medians) and distributions (using the Mann-Whitney U statistic) showed no statistically significant differences between the groups. Looking at Spearman rho correlation coefficients, the number of total hours completed was not significantly correlated with any of the Diversity dimensions.

SAT-Verbal was positively correlated with DV2 Knowledge of Diverse Perspectives (.171*). SAT-Math was negatively correlated with DV3 Examining Diversity, History, and Culture (-.211*). There were no significant correlations with ACT score and the Diversity dimensions. GPA was positively correlated with DV1 Factual Knowledge (.188**).

Comparisons between courses and assignment types

The assignments, instructional setting, and instructor type of the work products collected were varied. Some courses were taught by tenure-line faculty while others were taught by part time instructors or lecturers; there was a statistical difference in the scores for DV1 and DV2 between

the two instructor types. The work products collected from tenure-line faculty courses scored higher on DV2 Knowledge of Diverse Perspectives than did those collected from non-tenure-faculty-taught courses. However, work products collected from courses taught by part-time instructors or lecturers scored higher on DV1 Factual Knowledge. Many of the courses sampled were online classes; there was a statistical difference in the scores on dimensions DV4 Evaluating Claims and Theories about Diversity between the classroom-based and online class types, with the classroom-based course products scoring higher. Finally, there were statistical differences in the in-class-completed work products vs. the out-of-class assignments. The out-of-class work products scored significantly higher on DV1 Factual Knowledge, DV2 Knowledge of Diverse Perspectives, and DV3 Examining Diversity, History, and Culture. The differences in in-class vs. out-of-class could not be calculated for DV4 because it was deemed "not applicable" for the in-class assignments.

Interrater Reliability

For each pair of scorers, there were a number of independently-scored common papers so that interrater reliability could be assessed. Table 5.1 shows the reliability measures for Diversity.

Table 5.1 Interrater Reliability for Diversity

Dimension	Percent Agreement	Plus Percent Adjacent	Krippendorff's Alpha	Spearman's Rho
DV1 Factual Knowledge	33.9%	91.1%	.230	.225
DV2 Knowledge of Diverse Perspectives and Their Roots	44.7%	84.2%	.188	.227
DV3 Examining Diversity, History, and Culture	53.6%	60.7%	.541	.563**
DV4 Evaluating Claims and Theories about Diversity	32.0%	76.0%	.386	.379

**significant at the .001

Interrater reliability is a measure of the degree of agreement between scorers, and provides information about the trustworthiness of the data. It helps answer the question—would a different set of scorers at a different time arrive at the same conclusions? In practice, interrater reliability is enhanced over time through scorer discussion, as well as through improvements to the scoring rubric. Percent Agreement, Percent Agreement Plus Adjacent, and Krippendorf's Alpha measure scorer agreement. Spearman's Rho measures consistency between scorers. The UNCW benchmarks are .67 for Krippendorff's alpha and .7 for Spearman's rho. See Appendix B in the General Education Assessment Report 2012 for a more complete discussion of these statistics and the determination of benchmark levels.

Comparing the results of the reliability indices for this study to the benchmark of .67 for Krippendorff's alpha and .7 for Spearman's rho, there are no dimensions in the rubric that meet

these standards. One of the dimensions (DV3) came close. The percent agreement was greater than 50.0% in for DV3 as well. Looking at percent adjacent (that is, the scores that were within one level of each other), we find that three dimensions had greater than 70% of scores within one level of each other. Overall, these various measures of reliability illustrate randomness in agreement and indicate that additional norming activities are required.

Scorer Feedback on Rubric and Required Assumptions

Scorer opinion about the fit of the rubric dimensions to the assignments was mixed. Only for DV1 Factual Knowledge did all scorers agree that it fit the assignments, either with or without assumptions. For DV2 Knowledge of Diverse Perspectives and DV4 Evaluating Claims and Theories about Diversity, the majority of scorers found that these dimensions fit, either as-is or with assumptions (81% for each). However, most scorers that responded to the survey reported that DV3 Examining Diversity, History, and Culture did not fit the assignment, even with assumptions (54.5%).

In response to the open-ended questions about the Diversity rubric, all respondents described some issue about the fit between the assignment prompt and the rubric dimensions. Many comments mentioned that the assignment prompts did not ask for much of what the quality criteria detailed. One scorer wrote that it was necessary for "scorers to make the assignment fit."

Though no assumptions were needed for DV1 Factual Knowledge, two scorers mentioned that it was problematic that they did not know the breadth and depth of the knowledge that they could expect the student to have. Scorers pointed out that, for DV2 Knowledge of Diverse Perspectives, students were not explicitly asked to attend to diversity, history, and culture for the assignments, and so it was assumed that some interaction between those was implied in the student work. For DV4 Evaluating Claims and Theories about Diversity, and due to the limits required by the assignment prompt, scorers assumed that students would evaluate the author's claims (from their assigned readings) and not necessarily a theory.

When asked for possible improvements that could be made to the rubric, scorers had several suggestions. Many suggestions centered on clarifying the wording of the rubric and making the criteria more specific: for instance, replacing "complete argument" with "compelling or persuasive argument". Several scorers pointed out that that particular use of "complete" was troubling. Other problematic rubric terms listed included "identify" vs. "explain"; "basic" vs. "major"; and "substantial" vs. "important". One scorer suggested adding a dimension or expanding the quality criteria to include "analyzing evidence".

In addition to clarifying and refining the rubric, many scorers provided suggestions about aligning the rubric and assignment. Suggestions to this end included providing students with specific opportunities to meet the quality criteria of the rubric such as assignments that require

comparisons of diverse groups, and requiring multiple interpretations within student work. Finally, four scorers commented that having an instructor-authored sample work product would be helpful in knowing what a correct answer should be.

DISCUSSION

This was the second study using a diversity rubric. However, the rubric used was significantly modified to align with the student learning outcomes of Living in Our Diverse Society, a new general education component introduced with University Studies in Fall 2011. The median score on all four dimensions was 2. Table 5.2 shows the percent of work products scored at a level 2 or higher and the percent of work products scored at a level 3 or higher for each dimension. Level 2 is the benchmark for general education courses.

Table 5.2 Diversity Percent of Sample Scored at Least 2 and at Least 3

Dimension	% of Work Products Scored 2 or higher	% of Work Products Scored 3 or Higher
DV1 Factual Knowledge	85.5%	45.3%
DV2 Knowledge of Diverse Perspectives and Their Roots	71.9%	26.5%
DV3 Examining Diversity, History, and Culture	61.8%	13.8%
DV4 Evaluating Claims and Theories about Diversity	82.6%	31.8%

The classes from which the student work was selected included 100-, 200-, and 300-level courses. It is not surprising that scores were highest on DV1 Factual Knowledge. More surprising is the fact that scores were second highest on DV4 Evaluating Claims and Theories about Diversity, however only about half of the papers were scored on this dimension. There was no meaningful difference in the scores between 100-, 200-, and 300-level courses, or, as already noted, related to the number of credit hours completed by students. Although there were several demographic differences in scores on a few dimensions, the most notable is that females scored higher than males on three of the four dimensions. While this may be related to the diversity content itself, it could also be influenced by the rubric criteria, in which higher scores require increased levels of explanation (versus identification of elements). Previous results have shown females score higher on the written communication dimension Content Development. More study would be needed to provide insight into the reasons for these differences.

Instructors were not expected to provide work from an assignment that covered all learning outcomes. Nonetheless, a look at which of the four Living in Our Diverse Nation learning outcomes were addressed by the assignments that instructors viewed as aligning with the rubric is useful. This comparison could provide information for future action.

Table 5.3 Alignment of Assignments to LDN Learning Outcomes

Rubric Dimension	LDN Learning Outcome	Number of Assignments Aligned
DV1	LDN 1. Describe and explain various themes and issues relevant to the study of human diversity.	8
DV2	LDN 3. Demonstrate an understanding of social and cultural influences that shape perspectives of various social groups, while considering the consequences of advantage and disadvantage.	6
DV3	LDN 2. Analyze and interpret evidence of the influence of human diversity on the history and present culture of the United States.	4
DV4	LDN 4. Evaluate claims, arguments, and theories related to the ways in which diversity has shaped and continues to shape identity and experience in the U. S.	4

One assignment covered all four outcomes, and in the two courses that provided work from two journal assignments, all four outcomes were covered across the two assignments. Two of the remaining assignments covered three outcomes, and one assignment addressed only one dimension, DV1 Factual Knowledge.

The assignments themselves were a major factor in how well students performed. A number of scorers spoke to the fact that some assignment prompts did not explicitly ask students to address key aspects of the dimensions/SLOs, especially related to DV2 and DV4, and many scorers suggested providing specific opportunities to meet the quality criteria of the high levels on the scale (for instance by requiring comparisons of diverse groups and multiple interpretations of events or theories). DV3, and by extension LDN 2, has a number of issues associated with it: the least number of assignments addressing it as well as the lowest scores. Since the issues point to both the rubric and the SLOs that were addressed, recommendations should be considered in both areas.

The Diversity rubric and correlation table are located in the following appendices.

APPENDIX 5.A DIVERSITY RUBRIC

	Benchmark 1	Milestone 2	Milestone 3	Capstone 4	Score
Knowledge of Human Diversity					
DV1 Factual knowledge (LDN1)	Use some terminology surrounding diversity, but identifies few, if any, of the basic elements of an issue or theme regarding human diversity.	Identifies some of the basic elements of an issue or theme regarding human diversity. Description is incomplete or contains some inaccuracies or misconceptions.	Accurately explains the major elements of an issue or theme regarding human diversity.	Provides a comprehensive, detailed, and accurate discussion of an issue or theme regarding human diversity.	
DV2 Knowledge of diverse perspectives and their roots (LDN3)	Identifies some elements of the perspectives of a specific social group or groups, but does not demonstrate an awareness of societal or cultural influences on those perspectives.	Identifies some elements of the perspectives of a specific social group or groups and provides some explanation of how culture and society influenced (and continue to influence) those perspectives.	Explains the important aspects of the perspectives of a specific social group or groups and discusses how culture and society influenced (and continue to influence) those perspectives.	Discusses in detail the perspectives of a specific social group or groups and comprehensively examines how culture and society influenced (and continue to influence) those perspectives.	
Thinking Critically about Human Diversity *This SLO is assessed using the Critical Thinking VALUE Rubric dimensions PLUS the following two dimensions that elicit specific evidence related to human diversity.*					
DV3 Examining diversity, history, and culture (LDN 2)	Presents a narrow set of evidence that has been taken as factual without questioning to describe the influence of human diversity on the history and/or present culture of the United States.	Presents evidence that has been taken mainly as factual with little questioning to support a discussion of the influence of human diversity on the history and/or present culture of the United States.	Supplies substantial evidence that is relevant and has undergone some amount of inspection to support the examination of the influence of human diversity on the history and/or present culture of the United States.	Supplies comprehensive evidence that is relevant and thoroughly vetted to support the detailed examination of the influence of human diversity on the history and/or present culture of the United States.	
DV4 Evaluating claims and theories about diversity (LDN4)	Attempts to provide evidence that backs up or disputes a claim, argument or theory regarding the interplay between diversity, identity and experience, however evidence is inaccurate or unrelated.	Provides some accurate evidence that backs up or disputes a claim, argument or theory regarding the interplay between diversity, identity and experience. Argument is not complete, and other evidence may be inaccurate or unrelated.	Presents an evidence-based, accurate and substantially complete argument for or against a claim, argument or theory regarding the interplay between diversity, identity and experience. May acknowledge other viewpoint(s).	Presents an evidence-based, accurate and well-thought-out argument for or against a claim, argument or theory regarding the interplay between diversity, identity and experience. Acknowledges competing viewpoint(s).	

UNCW DIVERSITY RUBRIC

UNCW Learning Goal *Diversity*

Students will describe and examine the importance and implications of human diversity.

Diversity constitutes the knowledge, skills and attitudes necessary to examine the importance and implications of cultural and ethnic human differences. Diversity examines the significance of historical, political, social, racial, ethnic and cultural realities through critical thinking to understand and explain their implications in human endeavors.

University Studies Living in Our Diversity Nation component student learning outcomes:

The student will:

- LDN 1. Describe and explain various themes and issues relevant to the study of human diversity. [Foundational Knowledge, Diversity]
- LDN 2. Analyze and interpret evidence of the influence of human diversity on the history and present culture of the United States. [Information Literacy, Critical Thinking]
- LDN 3. Demonstrate an understanding of social and cultural influences that shape perspectives of various social groups, while considering the consequences of advantage and disadvantage. [Foundational Knowledge, Inquiry, Diversity]
- LDN 4. Evaluate claims, arguments, and theories related to the ways in which diversity has shaped and continues to shape identity and experience in the U. S. [Information Literacy, Critical Thinking, Diversity]

APPENDIX 5.B CORRELATIONS BETWEEN DIVERSITY DIMENSIONS

Spearman rho Rank Order Correlation Coefficients

Correlations

			GPA	TOTAL HOURS	ACT	SATV	SATM	DV1	DV2	DV3	DV4
Spearman's rho	GPA	Correlation Coefficient		-.199**	.418**	.385**	.248**	.188**	.067	.129	.145
		N		256	63	183	183	256	185	123	132
	TOTAL HOURS	Correlation Coefficient	-.199**		-.257*	-.160*	-.111	.034	-.011	-.039	-.086
		N	256		63	183	183	256	185	123	132
	ACT	Correlation Coefficient	.418**	-.257*		.710**	.585**	.057	-.029	-.126	-.042
		N	63	63		31	31	63	47	29	35
	SATV	Correlation Coefficient	.385**	-.160*	.710**		.384**	.083	.171*	-.107	.196
		N	183	183	31		183	183	135	88	99
	SATM	Correlation Coefficient	.248**	-.111	.585**	.384**		-.066	-.146	-.211*	-.107
		N	183	183	31	183		183	135	88	99
	DV1	Correlation Coefficient	.188**	.034	.057	.083	-.066		.707**	.657**	.631**
		N	256	256	63	183	183		185	123	132
	DV2	Correlation Coefficient	.067	-.011	-.029	.171*	-.146	.707**		.618**	.740**
		N	185	185	47	135	135	185		101	113
	DV3	Correlation Coefficient	.129	-.039	-.126	-.107	-.211*	.657**	.618**		.529**
		N	123	123	29	88	88	123	101		29
	DV4	Correlation Coefficient	.145	-.086	-.042	.196	-.107	.631**	.740**	.529**	
		N	132	132	35	99	99	132	113	29	

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

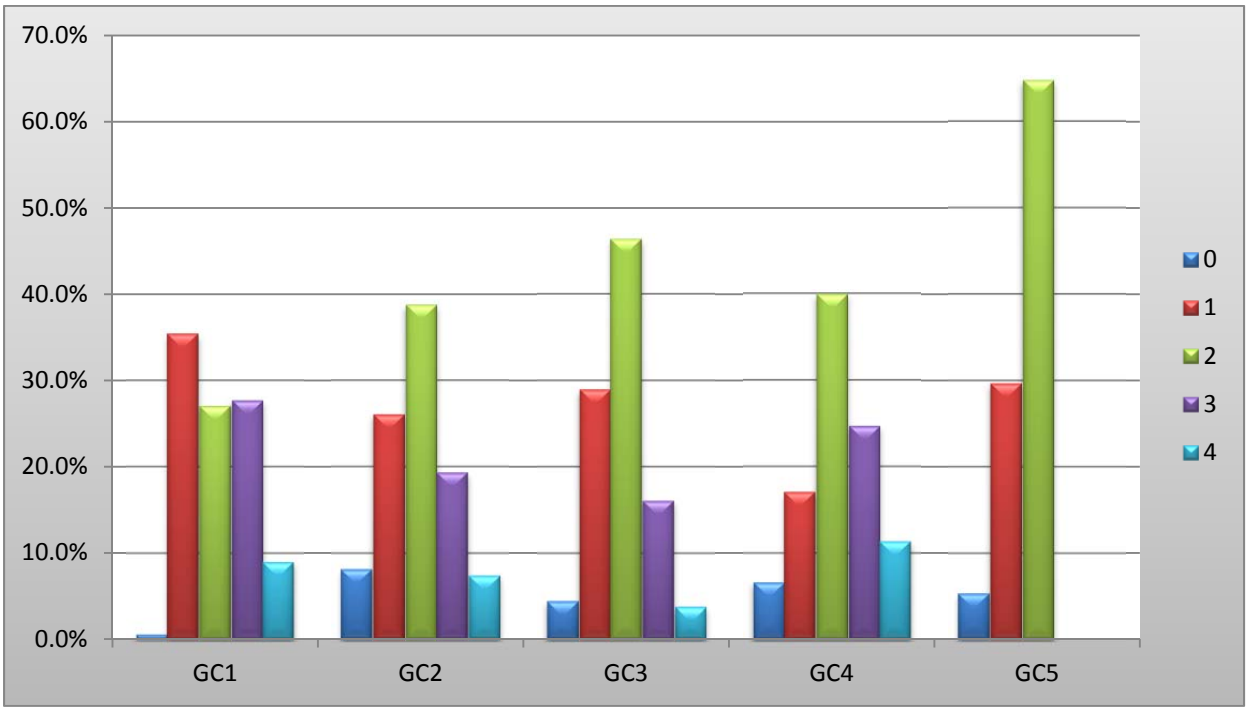
6. GLOBAL CITIZENSHIP

The UNCW Global Citizenship Learning Goal is for students to describe and examine the intellectual and ethical responsibilities of active global citizenship. For purposes of this Learning Goal, global citizenship is characterized by the ability to evaluate large-scale impacts of historical, scientific, economic, political cultural and artistic perspectives on individuals, societies and our environment; and by participation in efforts to make the world a better place (UNCW Learning Goals, 2011). The Global Citizenship rubric was developed in-house at UNCW and details five dimensions of knowledge related to human diversity—four that assess the Living in a Global Society component student learning outcomes and one that addresses ethically responsibility, which is part of the UNCW Global Citizenship Learning Goal. The Global Citizenship rubric can be found in Appendix 6.A at the end of this chapter. Four components of University Studies have at least one student learning outcome aligned to Global Citizenship. In this study, student work was sampled from the Living in a Global Society component.

SUMMARY OF SCORES BY DIMENSION

Eight faculty scorers scored 155 work products from six assignments across five courses from the Spring 2012 semester: ECN 326, INT 105, PAR 125, PLS 111, and SOC 240. A variety of assignment types were sampled; both in-class and out-of-class work, short and long written pieces, online discussion postings, blog entries, journal entries, and exam essay questions. Thirty-eight work products (24.5%) were scored by multiple scorers. Scorers determined that not all dimensions were applicable for all assignments. No one assignment was scored across-the-board on all five Global Citizenship dimensions. Two dimensions were aligned with all assignments. However, different scoring pairs deemed some dimensions applicable for an assignment while others did not. Figure 6.1 provides the score distributions for each dimension for work products that were scored on that dimension (i.e., work products scored as NA are not included).

GLOBAL CITIZENSHIP RESULTS BY DIMENSION FOR APPLICABLE SCORES ONLY



	GC1	GC2	GC3	GC4	GC5
0	0.6%	8.2%	4.5%	6.7%	5.4%
1	35.5%	26.1%	29.0%	17.1%	29.7%
2	27.1%	38.8%	46.5%	40.0%	64.9%
3	27.7%	19.4%	16.1%	24.8%	0.0%
4	9.0%	7.5%	3.9%	11.4%	0.0%
25th %tile	1	1	1	2	1
50th %tile	2	2	2	2	2
75th %tile	3	3	2	3	2
Mode	1	2	2	2	2
N	155	134	155	105	37

Figure 6.1 Distribution of Scores for Global Citizenship, Applicable Scores Only

RESULTS BY DIMENSION

GC1 Factual Knowledge

All work products were scored on this dimension. Scores on this dimension were the highest along with GC4 Tolerance of Differences. Fewer than one out of one hundred work products failed to show evidence of factual knowledge related to Global Citizenship. Just more than one-third of work samples were scored as identifying few facts associated with the relevant global issues, processes, trends, and systems, with some factual errors perhaps present (scores of 1). Slightly more than a quarter of the papers identified the main facts associated with global issues (scores of 2), and the same proportion identified the main facts *and* explained details of a few (scores of 3). About one in ten papers thoroughly discussed the main facts associated with the relevant global issues, processes, trends, and systems (scores of 4).

GC2 Knowledge of Connections within Systems

This dimension was scored for 86.5% of the sample work products collected. Scores on this dimension were in the middle range of scores for Diversity. One in ten work products failed to exhibit knowledge of the connections within global systems (scores of 0). About one-fourth of the samples collected described components of global systems without demonstrating understanding of the interconnectedness between global systems or processes (scores of 1). Almost two out of five work products demonstrated a limited understanding of the interconnectedness between global systems (scores of 2). One in five student work samples demonstrated a more complete, though still basic understanding of global systems (scores of 3). Less than one in ten products exhibited a nuanced understanding of the interconnectedness between and within global systems (scores of 4).

GC3 Use of Diverse Cultural Frames of References and Alternative Perspectives

This dimension was deemed applicable for all work products. The scores on this dimension were in the middle range of scores. One in twenty work products scored a zero on this dimension. Three out of ten student work samples considered only one's own perspective or used only one frame of reference when discussing global issues (scores of 1). Almost half of the samples showed evidence that the students acknowledged another perspective as well as one's own or used two frame of reference (scores of 2). One in six products both considered and applied multiple perspectives to the discussion of global issues (scores of 3). Fewer than four out of one hundred work samples demonstrated an on-going exploration and integration of multiple perspectives and/or frames of reference in addition to one's own when discussing global issues (scores of 4).

GC4 Tolerance of Differences

This dimension was viewed as applicable and was scored for 68.8% of the sample work. Scores on this dimension were among the highest of the dimensions scored. Fewer than one in ten work products failed to show a tolerance of differences when it was expected to be present (scores of 0). Around one in six work samples showed minimum acceptance of cultural differences (scores of 1). Two in five work products occasionally evidenced acceptance of cultural differences, though may have been trouble by ambiguous situations, perhaps with a fixed idea about what "should" occur (scores of 2). One quarter of the work products showed acceptance of obvious cultural differences, and showed some flexibility about what "should" occur based on that acceptance (scores of 3). Just over one in ten samples demonstrated acceptance of cultural differences, including subtle or hidden differences (scores of 4).

GC5 Ethical Responsibility

This dimension was scored for 23.9% of the student work products collected for Global Citizenship. Scores for this dimension were in the lower range, with no scores of three or four. One in twenty work products scored a zero for GC5, indicating that there was no evidence of

knowledge of ethical responsibility, though it would be appropriate for the given assignment. Three out of ten work products described events which have changed or could change global society or environment with little or no mention of ethical details (scores of 1). Almost two-thirds of the products scored for GC5 identified ethical dimensions of particular acts and decisions that either have changed or could change global society or environment (scores of 2). Again, no scores of three or four were given on this dimension.

Correlation between Dimensions

Many of the dimension scores were correlated with each other at the .01 or .05 level of significance, with the highest correlation between GC1 Factual Knowledge and GC2 Knowledge of Connections between Systems. GC5 Ethical Responsibility was not correlated with GC1, GC2, or GC3. While a few correlation coefficients were high, overall they do not exhibit a pattern that might suggest scorers were not scoring each dimension separately. See Appendix 6.B for a complete presentation of correlation coefficients.

Demographic and Preparedness Findings

There were no statistically significant difference between the means, medians, and the score distributions of male and female students, transfer vs. UNCW-start students, and honors vs. non-honors students. The samples of students with race/ethnicities other than white were too small to compare the groups.

To compare scores based on number of credit hours completed, two methods were used. First, students were grouped into four categories, those having completed 0 – 30 credit hours (21.9% of the sample), 31 – 60 credit hours (34.8% of the sample), 61 – 90 (17.4% of the sample), and over 90 credit hours (25.8% of the sample). Comparison of means (using ANOVA), medians (using Independent Samples test of medians) and distributions (using the Mann-Whitney U statistic) showed statistically significant differences between the groups for GC2 Knowledge of Connections within Systems and GC3 Use of Diverse Cultural Frames of Reference and Alternative Perspectives. For both dimensions, seniors were the lowest-scoring group, while freshman scored the highest for GC2 and sophomores the highest for GC3. Looking at Spearman rho correlation coefficients, the number of total hours completed was negatively correlated with GC2 (-.239**) and GC3 (-.159*).

There were no significant correlations with ACT and SAT-Math scores and the Global Citizenship dimensions. SAT-Verbal was positively correlated with GC5 Ethical Responsibility (.554**).

COMPARISONS BETWEEN COURSES AND ASSIGNMENT TYPES

The assignments, instructional setting, and instructor type of the work products collected were varied. However, the numbers of each type collected and subsequently scored for particular variables were not large enough for statistical comparison.

INTERRATER RELIABILITY

There were a number of common student work products in each scorer's packet by scoring pair so that interrater reliability could be assessed. Table 6.1 shows the reliability measures for Global Citizenship.

Table 6.1 Interrater Reliability for Global Citizenship

Dimension	Percent Agreement	Plus Percent Adjacent	Krippendorff's Alpha	Spearman's Rho
GC1 Factual Knowledge	38.5%	84.6%	.270	.342*
GC2 Knowledge of Connections within Systems	58.9%	88.2%	.372	.432*
GC3 Use of Diverse Cultural Frames of References and Alternative Perspectives	48.7%	89.7%	.193	.279
GC4 Tolerance of Differences	38.7%	71.0%	.191	.222
GC5 Ethical Responsibility	42.9%	71.4%	.348	.575

Interrater reliability is a measure of the degree of agreement between scorers, and provides information about the trustworthiness of the data. It helps answer the question—would a different set of scorers at a different time arrive at the same conclusions? In practice, interrater reliability is enhanced over time through scorer discussion, as well as through improvements to the scoring rubric. Percent Agreement, Percent Agreement Plus Adjacent, and Krippendorf's Alpha measure scorer agreement. Spearman's Rho measures consistency between scorers. The UNCW benchmarks are .67 for Krippendorff's alpha and .7 for Spearman's rho. See Appendix B in the General Education Assessment Report 2012 for a more complete discussion of these statistics and the determination of benchmark levels.

Comparing the results of the reliability indices for this study to the benchmark of .67 for Krippendorff's alpha and .7 for Spearman's rho, there are no dimensions in the rubric that meet these standards. Looking at percent adjacent (that is, the scores that were within one level of each other), we find that all dimensions had greater than 70% of scores within one level of each other.

Overall, these various measures of reliability illustrate randomness in agreement and indicate that additional norming activities are required.

Scorer Feedback on Rubric and Required Assumptions

Scorer opinion of the fit of the rubric dimensions to the assignments was somewhat mixed. Scorers generally felt that GC1 Factual Knowledge, GC2 Knowledge of Connections within Systems, and GC3 Use of Diverse Cultural Frames of Reference fit the assignment, either as-is or with assumptions (only two scorers felt that GC2 did fit at all). However, for GC4 Tolerance of Differences and GC5 Ethical Responsibility, many scorers felt that these dimensions did not fit. This is likely because only a few assignments were deemed as fitting these dimensions, and scorers were not exposed to the gamut of assignments.

Most of the open-ended feedback about the fit of dimensions to assignment was about the limitations of the assignments themselves. For example, one scorer commented, "it was assumed that discussing multiple perspectives was implied in the assignment instructions" since it was not explicitly required. Additional comments along these same lines were about assuming the use of multiple frames of reference, that students selected a global issue when it was not specified to do so, and that students would present their own opinions in their writing, though the assignments did not necessarily ask for this.

Many scorers commented that the assignment limited the level to which student could achieve. Suggestions for improving the assignment included emphasizing the "global aspect of the assignment", to be clear in the directions about requiring multiple viewpoints and to investigate similarities and differences in culture groups, and to provide more on-going guidance from the instructor as students go through the writing process.

With regards to the rubric quality criteria, several scorers mentioned that ethical responsibility and tolerance of differences were not appropriate criteria for all assignments, especially research papers. There were several suggestions that specific words in the rubric be defined, such as "global systems", "global", "critical thinking", and "multiple perspectives".

Discussion

This was the first use of the Global Citizenship rubric. The rubric used was written to align with both the Global Citizenship Learning Goal and student learning outcomes of Living in a Global Society, a new general education component introduced with University Studies in Fall 2011. The median score on all four dimensions was 2. Table 6.2 shows the percent of work products scored at a level 2 or higher and the percent of work products scored at a level 3 or higher for each dimension.

Table 6.2 Global Citizenship Percent of Sample Scored at Least 2 and at Least 3

Dimension	% of Work Products Scored 2 or higher	% of Work Products Scored 3 or Higher
GC1 Factual Knowledge	63.9%	36.8%
GC2 Knowledge of Connections	65.7%	26.9%
GC3 Use of Diverse Cultural Frames	66.5%	20.0%
GC4 Tolerance of Differences	76.2%	36.2%
GC5 Ethical Responsibility	64.9%	0.0%

The classes from which the student work was selected included 100-, 200-, and 300-level courses. The fact that two in three students (three in four for GC4) met the Universities Studies benchmark is a good indication for this new component of the general education curriculum. However, the fact that the scores are at best not improving with credit hours completed, and at worse negatively correlated for GC2 and GC3, indicates that more emphasis of Global Citizenship is needed either in University Studies or the majors.

Instructors were not expected to provide work from an assignment that covered all learning outcomes. Nonetheless, a look at which of the four Living in a Global Society learning outcomes were addressed by the assignments that instructors viewed as aligning with the rubric is useful. This comparison could provide information for future action.

Table 6.3 Alignment of Assignments to LGS Learning Outcomes and UNCW Learning Goal Global Citizenship

Rubric Dimension	LGS Learning Outcome	Number of Assignments Aligned
GC1	GS 1. Demonstrate knowledge of global issues, processes, trends, and systems.	6
GC2	This dimension aligns with, but is an extension of, GS 1, as it addresses students' understanding of connections within and among systems.	5
GC3	GS 2. Use knowledge, diverse cultural frames of reference, and alternate perspectives to think critically and solve problems.	6
GC4	GS 3. Accept cultural differences and tolerate cultural ambiguity.	6
GC5	Global Citizenship: Students will describe and examine the intellectual and ethical responsibilities of active global citizenship.	3

The coverage of the Living in Our Global Society (GS) learning outcomes was quite good, with all assignments aligning to the three GS outcomes. With less than a quarter of the papers scored on Ethical Responsibly, and the fact that ethical responsibility is not directly addressed in the Living in a Global Society student learning outcomes, an examination of the relationship between the curriculum and the UNCW Learning Goal is warranted.
The Global Citizenship rubric and correlation table are located in the following appendices

APPENDIX 6.A GLOBAL CITIZENSHIP RUBRIC

	Benchmark 1	Milestone 2	Milestone 3	Capstone 4	Score
GC1 Factual knowledge	Identifies few facts associated with the relevant global issues, processes, trends, and systems; factual errors may be present.	Identifies the main facts associated with the relevant global issues, processes, trends, and systems.	Identifies the main facts associated with the relevant global issues, processes, trends, and systems, explaining the details of a few.	Thoroughly discusses the main facts associated with the relevant global issues, processes, trends, and systems.	
GC2 Knowledge of connections within systems	Describes components of global system(s) without demonstrating understanding of the interconnectedness within global system(s) and process(es).	Demonstrates a limited understanding of the interconnectedness within global system(s) and process(es).	Demonstrates a fairly complete, yet basic, understanding of the interconnectedness within global system(s) and process(es).	Demonstrates a nuanced understanding of the interconnectedness within complex global system(s) and process(es).	
Thinking Critically and Solving Problems This SLO is assessed using the Critical Thinking or Problem Solving VALUE Rubric dimensions **PLUS** **GC3 Use of diverse cultural frames of reference and alternative perspectives**	Considers only one's own perspective or uses only one frame of reference when discussing global issues.	Acknowledges another perspective as well as own or uses at least two frames of reference when discussing global issues.	Considers and applies multiple perspectives and/or frames of reference in addition to one's own when discussing global issues.	Demonstrates on-going exploration and integration of multiple perspectives and/or frames of reference in addition to one's own when discussing global issues.	
GC4 Tolerance of differences	Shows minimum acceptance of cultural differences and evidences a self-centric view of culture.	Occasionally shows acceptance of cultural differences, though may be troubled by ambiguous situations. May have fixed ideas about what "should" occur.	Shows acceptance of obvious cultural differences, and can manage in some ambiguous situations. Shows some flexibility about what "should" occur based on that acceptance.	Shows acceptance of cultural differences, including subtle or hidden differences, and is not troubled by ambiguous situations. Has flexible ideas about what "should" occur even in complex situations.	
GC5 Ethical responsibility	Describes events which have changed or could change global society or environment with little or no mention of ethical dimensions.	Identifies ethical dimensions of particular acts and decisions that either have changed or could change global society or environment.	Considers the ethical dimensions of own acts and decisions that could either enhance or diminish global society or environment.	Demonstrates commitment to act, live, and create ethically in an attempt to enhance global society or environment.	

UNCW GLOBAL CITIZENSHIP RUBRIC

UNCW Learning Goal *Global Citizenship*

Describe and examine the intellectual and ethical responsibilities of active global citizenship

Global citizenship is characterized by the ability to evaluate large-scale impacts of historical, scientific, economic, political cultural and artistic perspectives on individuals, societies and our environment; and by participation in efforts to make the world a better place. (UNCW Learning Goals Definitions)

University Studies: Living in a Global Society component Student Learning Outcomes:

The student will:

- GS 1. Demonstrate knowledge of global issues, processes, trends, and systems. [Foundational Knowledge]
- GS 2. Use knowledge, diverse cultural frames of reference, and alternate perspectives to think critically and solve problems. [Foundational Knowledge, Inquiry, Critical Thinking, Diversity, Global Citizenship]
- GS 3. Accept cultural differences and tolerate cultural ambiguity. [Global Citizenship]

References used in creating rubric:

Adelphi University General Education Learning Goal Rubric – Global Citizenship http://academics.adelphi.edu/gened/goals.php

University of Southern California Assessment Criteria – Social/Behavior Sciences
http://www.ipr.sc.edu/effectiveness/criteria/socibeh.htm

Washington State University Global Outcomes and Rubrics
http://ip.wsu.edu/education_abroad/outcomes-program-assessment/education-abroad-learning-outcomes.html

APPENDIX 6.B CORRELATIONS BETWEEN GLOBAL CITIZENSHIP DIMENSIONS

Spearman rho Rank Order Correlation Coefficients

			GPA	TOTAL HOURS	ACT	SATV	SATM	GC1	GC2	GC3	GC4	GC5
Spearman's rho	GPA	Correlation Coefficient		.241**	.314	.246**	.245**	.123	-.012	.121	-.046	-.004
		N		154	30	114	114	154	133	154	105	37
	TOTAL HOURS	Correlation Coefficient	.241**		.083	.005	.073	-.142	-.239**	-.159*	-.119	-.061
		N	154		30	114	114	155	134	155	105	37
	ACT	Correlation Coefficient	.314	.083		.587*	.166	.271	.282	.221	-.018	.296
		N	30	30		17	17	30	25	30	19	5
	SATV	Correlation Coefficient	.246**	.005	.587*		.236*	-.031	.053	.074	.043	.554**
		N	114	114	17		114	114	100	114	76	29
	SATM	Correlation Coefficient	.245**	.073	.166	.236*		-.042	-.065	.027	-.097	.076
		N	114	114	17	114		114	100	114	76	29
	GC1	Correlation Coefficient	.123	-.142	.271	-.031	-.042		.683**	.500**	.397**	.202
		N	154	155	30	114	114		134	155	105	37
	GC2	Correlation Coefficient	-.012	-.239**	.282	.053	-.065	.683**		.541**	.399**	.390*
		N	133	134	25	100	100	134		134	84	37
	GC3	Correlation Coefficient	.121	-.159*	.221	.074	.027	.500**	.541**		.460**	.304
		N	154	155	30	114	114	155	134		105	37
	GC4	Correlation Coefficient	-.046	-.119	-.018	.043	-.097	.397**	.399**	.460**		-.068
		N	105	105	19	76	76	105	84	105		37
	GC5	Correlation Coefficient	-.004	-.061	.296	.554**	.076	.202	.390*	.304	-.068	
		N	37	37	5	29	29	37	37	37	37	

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

7. General Discussion and Recommendations

This chapter provides a general discussion across all studies. It includes an overall discussion of findings regarding student abilities, scorer and instructor feedback on the process, an overall discussion of interrater reliability, follow up on prior-year recommendations, including a study on actions taken by faculty, and new recommendations.

UNCW Student Abilities on Learning Goals

Five of the eight UNCW Learning Goals were assessed in 2011-2012 using course-embedded assignments. Table 7.1 combines the results from all studies, and presents the % of work products that met or exceeded the performance benchmark, or proficiency level, for each dimension of each rubric.

Table 7.1 Percent of Student Work Products Meeting Performance Benchmarks

Courses	Dimension	% of Work Products Scored at Proficiency Level or Higher
MAT 151	FK Math Linear Equations	91.7%
	FK Math Exponential Equations	61.5%
BIOL 105 CHM 101	IN1 Topic Selection	NA
	IN2 Existing Knowledge	NA
	IN3 Design Process	87.3%
	IN4 Analysis	72.6%
	IN5 Conclusions	77.0%
	IN6 Limitations and Implications	39.5%
FRH 201	SL WR1 French Mechanical and Spelling	71.0%
	SL WR2 French Grammar	63.3%
	SL WR3 French Following Instructions	77.8%
	SL WR4 French Content, Vocabulary & Style	65.7%
SPN 201	SL WR1 Spanish Content	90.3%
	SL WR2 Spanish Organization	83.0%
	SL WR3 Spanish Vocab. – appropriateness & variety	52.1%
	SL WR4 Spanish Vocab. – proper use	76.7%
	SL WR5 Spanish Grammar	82.0%
	SL RD Spanish	82.6%
GRN 101, HST 204 MUS 117, PSY 270 PSY 271, SOC 325	DV1 Factual Knowledge	85.5%
	DV2 Knowledge of Diverse Perspectives and Their Roots	71.9%
	DV3 Examining Diversity, History, and Culture	61.8%
	DV4 Evaluating Claims and Theories about Diversity	82.6%
ECN 326, INT 105 PAR 125, PLS 111 SOC 240	GC1 Factual Knowledge	63.9%
	GC2 Knowledge of Connections	65.7%
	GC3 Use of Diverse Cultural Frames	66.5%
	GC4 Tolerance of Differences	76.2%
	GC5 Ethical Responsibility	64.9%

For eight dimensions (29.6%), 80% or more of the student work met the benchmark, and for 15 (55.5%), 70% or more of the student work met the benchmark. The highest levels of work meeting the benchmarks were in FK Math Linear Equations and SL WR1 Spanish Content. Comparing the results within each rubric indicates specific areas that need to additional coverage and opportunities for practice.

Only two dimensions stand out with less than 60% of work meeting the benchmark: IN6 Limitations and Implications, with only 39.5% at or above proficiency, and SL WR3 Spanish Vocabulary – appropriateness and variety, with 52.1% at or above proficiency. The first of these, IN6 Limitations and Implications, has been assessed in the past. Assessing student work from ENG 201 and PSY 105 in 2010, 62.5% of the work met the benchmark. In pilot study in a major in 2011, only 14.3% of the work met the level 2 benchmark, and no papers were scored a 4, and it was clear to the faculty scorers that this dimension was overlooked in the assignment instructions. These findings point towards more focus on this dimension of Inquiry. For the second of these dimensions, SL WR3 Spanish Vocabulary – appropriateness and variety, departmental discussions about appropriate vocabulary usage seem to be an fitting course of action.

Scorer Feedback on Process

Table 7.2 provides the combined results for the survey items for the Inquiry, Second Language, Global Citizenship, and Diversity scoring processes.

Table 7.2 Scorer Feedback on Process

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Blank
The invitation to volunteer accurately described the experience.	0 (0%)	0 (0%)	2 (5.6%)	5 (13.9%)	29 (80.6%)	0 (0%)
The timing of the invitation gave adequate opportunity for attending workshops and scoring.	0 (0%)	0 (0%)	0 (0%)	4 (11.1%)	32 (88.9%)	0 (0%)
The norming session adequately prepared me for what was expected of me during the scoring session.	0 (0%)	0 (0%)	1 (2.8%)	12 (33.3%)	22 (61.1%)	1 (2.8%)
The scoring session was well-organized.*	0 (0%)	0 (0%)	0 (0%)	3 (10.7%)	25 (89.3%)	0 (0%)
The structure of the scoring made it reasonable to work for the full time.	0 (0%)	0 (0%)	1 (2.8%)	9 (25.0%)	26 (72.2%)	0 (0%)
When I had questions, one of the leaders was available to answer it.	0 (0%)	0 (0%)	0 (0%)	1 (2.8%)	35 (97.2%)	0 (0%)
When I had questions, the question was answered.	0 (0%)	0 (0%)	0 (0%)	4 (11.1%)	32 (88.9%)	0 (0%)
I was comfortable scoring student work products from outside my discipline on the broad Learning Goals.**	0 (0%)	1 (6.3%)	2 (12.5%)	11 (68.8%)	2 (12.5%)	0 (0%)
The process is an appropriate way to assess students on the UNCW Learning Goals.	1 (2.8%)	3 (8.3%)	4 (11.1%)	18 (50.0%)	10 (27.8%)	0 (0%)
This process is valuable in improving student learning.	0 (0%)	0 (0%)	14 (38.9%)	10 (27.8%)	12 (33.3%)	0 (0%)
I would participate in this process again.	0 (0%)	0 (0%)	2 (5.6%)	8 (22.2%)	26 (72.2%)	0 (0%)
I would recommend participating in this process to my colleagues.	0 (0%)	0 (0%)	3 (8.3%)	5 (13.9%)	28 (77.8%)	0 (0%)

*Only on the surveys for Inquiry, Global Citizenship, and Diversity scorers

**Only on the surveys for Global Citizenship and Diversity scorers

There were also three open-ended questions on the survey. The results of both the Likert-scale and open-ended questions are discussed below.

There was a high level of satisfaction with regard to most aspects of the process. The initial contact and explanations about the responsibilities of the volunteers was clear to most (two scorers responded neutrally). Most scorers were positive in their responses about the norming session, for example, responding that the "process went smoothly and expectations were clear."

However, one scorer responded neutrally that the norming session prepared him well for the scoring. Five of the scorers noted, in the open-ended questions, that the norming session was the best part of the scoring process in general. Positive comments were that it was interesting, clear, and well-organized, that it was helpful, and that it made scorers more comfortable about the reliability of the assessment. One scorer mentioned that participating in the norming and scoring process led to self-reflection on personal teaching practices. Another scorer mentioned that the process was "good training for teachers that teach those [science] labs".

Regarding the scoring session question, the Second Language scoring took place individually and those scorers did not meet for a scoring session. Therefore, these scorers did not answer the survey question about the scoring session in particular. All other respondents (for Inquiry, Global Citizenship, and Diversity) felt that the scoring session was well-organized and allowed enough time for scoring. Additionally, all of the scorers were able to ask questions and have them answered by one of the scoring session leaders. Two respondents mentioned, in the open-ended questions, how smoothly the process went, and appreciated the clarity of instruction. Two scorers appreciated being able to work with a partner. One person mentioned that the rubric was easy to follow.

Although scorers felt prepared to score, they did provide recommendations for improving preparation. One person noted that it might be helpful to increase the number of papers scored during the norming session. Two scorers suggested that a longer training session would be useful. Another scorer suggested splitting the norming session into two sessions to allow scorers time to process their thoughts about the rubric. Half of the scorers mentioned the need to increase the alignment of the assignment to the rubric as a means of improving the process as a whole. Another person thought that student writing ability should also be scored when scoring for Inquiry. One scorer wrote that half-scores would be useful for scoring student work. One scorer from a group that scored on their own suggested an alternative means of securing student work to minimize paper mix-ups. Finally, one scorer suggested an in-depth look into the meaning of some of the qualifying words on the rubric, such as what is meant by "significant" would be helpful.

Looking at the broader survey questions, most scorers for which this question was applicable (76.5%) felt comfortable scoring outside of their discipline (only Global Citizenship and Diversity scorers scored work outside of their discipline; the Inquiry and Second Language scorers were from the departments from which the student work was collected). Over 75% of scorers agreed that the process was an appropriate way to assess students on the UNCW learning goals, although 11% responded neutrally to this question, and another 11% disagreed. Most scorers that disagreed appear to have questions about the assignment rather than the rubric or process. Over half (61.1%) of the scorers agreed that the process is valuable in improving student learning, and 38.9% of the scorers were neutral on the same topic. One scorer noted that

participating in the process was very useful to her, as an instructor in the discipline. Most scorers would recommend participating in the process to their colleagues and would do it again themselves (three responded neutrally).

Instructor Feedback

A brief survey was sent to the 12 instructors who provided the student work products. Five responded. This is similar to previous years' response rates. Instructors were asked to comment on the process of selecting an assignment, the collection process, and any other parts of the process.

Four respondents said that the assignment selection process was not difficult, and that they felt their assignment fit the rubric well. One mentioned that, in some sense, every assignment in the course would have been appropriate because the course was designed around the learning goal. Another mentioned that since the course objectives were aligned with the Diversity learning goal, that there were already several assignments in place from which to choose. Similar comments were made by the other two faculty members who indicated that the selection process was not difficult. These comments are at odds with scorer comments about the fit of the rubrics to the assignments. This discrepancy needs to be studied and remedied. One respondent noted the assignment selection process as being very difficult, mentioning that since most assessments are designed to demonstrate student knowledge at a single point in time and not as change over time, that the only empirical way to demonstrate student learning would be with a pre- and post-test of student knowledge.

Regarding the collection process, three respondents said that the process went very smoothly. Within these responses were instructors who collected the work products in hard copy as well as one instructor whose course was taught in the Blackboard system. One faculty member mentioned that their graduate assistants were somewhat confused by the process, and that having them meet with the assessment office would have be helpful.

When asked for additional comments on the experience, one instructor underscored his concern about not using a pre/post test methodology for his course. The faculty members rated the entire experience between 3 and 5 on a 5-point scale, and one mentioned that it was "easier than expected" while another commented that it was not onerous, but was extra work.

Interrater Reliability

Table 7.3 combines the interrater reliability (IRR) findings from all 2011-2012 studies.

Table 7.3 Interrater Reliability

Dimension	Percent Agreement	Plus Percent Adjacent	Krippendorff's Alpha	Spearman's Rho
IN6 Limitations and Implications	73.3%	93.3%	.870	.859**
SPN Reading	82.4%	98.4%	.765	.772**
FRH-W1 Mechanical and Spelling	46.7%	60.0%	.697	.751**
FRH-W2 Grammar	40.0%	86.7%	.662	.817**
FRH-W3 Following Instructions	46.7%	73.3%	.660	.716**
IN5 Conclusions	69.3%	92.0%	.617	.619**
DV3 Examining Diversity, History, and Culture	53.6%	60.7%	.541	.563**
IN3 Design Process	61.3%	97.3%	.539	.531**
FRH-W4 Sophistication of Writing and Communication of Information	40.0%	80.0%	.502	.769**
SPN-W5 Grammar	36.8%	84.2%	.424	.458*
DV4 Evaluating Claims and Theories about Diversity	32.0%	76.0%	.386	.379
SPN-W2 Organization	42.1%	78.9%	.381	.441
GC2 Knowledge of Connections within Systems	58.9%	88.2%	.372	.432*
IN4 Analysis	48.0%	93.3%	.359	.358**
GC5 Ethical Responsibility	42.9%	71.4%	.348	.575
GC1 Factual Knowledge	38.5%	84.6%	.270	.342*
DV1 Factual Knowledge	33.9%	91.1%	.230	.225
GC3 Use of Diverse Cultural Frames of References and Alternative Perspectives	48.7%	89.7%	.193	.279
GC4 Tolerance of Differences	38.7%	71.0%	.191	.222
DV2 Knowledge of Diverse Perspectives and Their Roots	44.7%	84.2%	.188	.227
SPN-W1 Content	31.6%	78.9%	.128	.291
SPN-W3 Vocabulary: appropriateness and variety of lexical terms and idiomatic expressions	26.3%	57.9%	.070	.276
SPN-W4 Vocabulary: proper use	36.8%	78.9%	.057	.090

The highest IRR results were for the French Writing and Inquiry rubrics. On the French rubric, all four dimensions met the Spearman benchmark, and 3 of 4 met the Krippendorf agreement benchmark. On the Inquiry rubric, one dimension met the Krippendorf alpha and Spearman

benchmarks, and all dimensions had high percent adjacent findings. Interrater reliability for Spanish Reading, which was a simple incorrect, partially correct, and correct rubric was very good. Besides the simplicity of the rubric, all scorers worked out the content requirements for correct and partially correct during the scoring workshop. Inquiry IRR was good, and Diversity and Global Citizenship reliability statistics were mediocre to good. Spanish Writing interrater reliability fared the lowest. Given that this rubric has been used the longest, the results were surprising. Work needs to continue with regard to interrater reliability, which will require enhancements to the scorer workshops, scoring events, and a review of, and possible modifications to, the rubrics.

Follow up on Previous Recommendations

The following explains the progress made on the recommendations from last year and those from the previous year that had not been completed by the time of the last report.

2010 Recommendations

Two of the five recommendations from 2010 were completed in 2011. Two require on-going efforts, and one was completed in 2012.

- *Levels of expected performance at the basic studies, or lower division, level should be developed for each rubric.*

With the implementation of additional components of University Studies that contain courses at the 300- and 400-level, it iss clear that there cannot be one benchmark for all University Studies courses. However, the minimum expectation for 100- and 200-level courses was set at level 2, and annual reports will continue to include the percent of work products scored at or above both 2 and 3. Levels 3 and 4 are appropriate standards of comparison for 300- and 400-level courses, respectively, and level 4 remains the benchmark for graduating seniors.

- *Additional exposure to the content of and rationale for the UNCW Learning Goals should be provided to increase faculty ownership and awareness of these Goals.*

This recommendation has been merged with a similar recommendation in 2011, which provided more specifics. See below.

- *Modifications and improvements to the general education assessment process should be made as needed, including the following: modify rubrics based on feedback, develop benchmarks work products, and enhance instructor and scorer workshops.*

Modifications continued in AY 2011-2012. Scoring workshops were increased in length; the Information Literacy rubric was modified slightly; and communication with instructors was enhanced.

2011 Recommendations

- *Continue efforts to improve interrater reliability.*

Slightly longer workshops for faculty scorers were implemented for most scoring sessions. Some scoring sessions were held in the all-day format, although two were half-day. In addition, the second language scoring was performed individually after a longer-format workshop. Interrater reliability results for Inquiry and Second Language reflect these efforts. Interrater reliability for Diversity and Global Citizenship were low despite these efforts, however the rubrics used were both newly-created.

- *Present targeted content workshops where we invite on-campus experts in various disciplines to share some of the specific ways they are introducing the UNCW Learning Goals into their courses*

This recommendation is an offshoot of another from 2010. Two workshops were presented in the 2011-2012 academic year: Global Citizenship and Thoughtful Expression. An Inquiry workshop is scheduled.

- *Introduce Learning Outcomes and guidelines for Explorations Beyond the Classroom and Thematic Transdisciplinary Clusters which will include a synthesis of information and aspects related to critical thinking. This will increase exposure to these learning goals.*

Learning outcomes and guidelines were created for Explorations beyond the Classroom. Transdisciplinary Thematic Clusters were launched with guidelines regarding learning outcomes, which include the requirement for at least one synthesis outcome for the cluster.

- *Develop a central information site where we can share summaries of reports, best practices of our colleagues, relevant literature and information from other institutions, and other information than may be helpful for faculty seeking to improve how they present information related to our learning goals. For this year, the goal would be to start this effort, with expansion in future years.*

The following websites were launched or improved:
http://www.uncw.edu/assessment/ The overall portal to assessment information at UNCW was updated.
http://www.uncw.edu/assessment/general/index.html A general education assessment website was launched containing information about the process for course faculty, students, and scorers, frequently asked questions, and findings presented by learning goal as well as complete annual reports.
http://www.uncw.edu/assessment/resources.html A site containing assessment and teaching resources was launched. Assessment resources are categorized by department. For the teaching resources, users can select to browse by department or by UNCW Learning Goal.

- *Reemphasize the desirability of including the relevant learning outcomes on syllabi of courses approved for the new University Studies curricula (or provide links to these goals on the syllabi).*

The University Studies Advisory Committee continues to strongly recommend that component student learning outcomes be listed and aligned to course SLOs on syllabi for all University Studies courses.

Actions Taken by Faculty

In 2010-2011, three workshops were given related to the Learning Goals of Critical Thinking, Information Literacy, and Thoughtful Expression. Workshops on Critical Thinking and Information Literacy were the first two workshops in the UNCW Learning Goals series. In addition, the University Studies Advisory Committee held a workshop for faculty writing proposals for the Writing Intensive and Information Literacy components of University Studies. This workshop was designed to help them align their course curriculum with the component student learning outcomes, and develop opportunities for learning and means of assessment.

In September, 2011, a survey was sent to participants to find out whether they had made changes to their courses after attending the workshops. A total of 46 responses were received, 25 from the Critical Thinking workshop (CT), 6 from the Information Literacy workshop (IL), and 15 from the Writing Intensive/Information Literacy proposal workshop (WI/IL). The overall response rate was 75.4%, and the response rates across the three workshops were 92.6% for CT, 60.0% for IL, and 62.5% for WI/IL. Table 7.4 shows the percent of respondents that reported that they had made changes, planned to make changes, or had no plans for making changes to their course(s) after attending the workshop.

Table 7.4 Workshop attendees' survey responses

Workshop	Made Changes	Plan to Make Changes	Do Not Plan to Make Changes
LG Series: Critical Thinking	12 (48.0%)	4 (16.0%)	9 (36.0%)
LG Series: Information Literacy	3 (50%)	1 (16.7%)	2 (33.3%)
University Studies: Writing Intensive and Information Literacy	5 (33.3%)	5 (33.3%)	5 (33.3%)
Overall	20 (43.5%)	10 (21.7%)	16 (34.8%)

The changes that were made or were planned to be made affect courses in the following disciplines:

Table 7.5 Courses in which Changes Were or Will Be Made

Critical Thinking	Information Literacy (from both workshops)	Written Communication
Biology (2) English (6) Nursing (5) Secondary Education Social Work (2) Spanish (2) First Year Seminars (3)	Biology English (2) Honors Music (3) Recreation, Sports Leadership, and Tourism Management Secondary Education Sociology (2) Spanish (2)	Chemistry Computer Science (2) French (2) Music (2) Sociology (2)

Participants were asked to explain the types of changes they made or planned to make. It was gleaned that some of respondents had already determined to make changes before attending the workshops, and that the workshops were helpful to those efforts. For many responses, there was a clear link between workshop material and changes made. For others, the changes seemed to be an intellectual outgrowth of thinking based on the workshop and other investigation.

There was variety to the types of changes reported for Critical Thinking. Five respondents reported that they are incorporating more critical thinking, open-ended, or what one person called “fundamental and powerful” questions into their class(es). Two reported the introduction of case studies. A number of the Nursing faculty explained that clinical experiences were being enhanced with an emphasis on critical thinking. Three respondents reported presenting lessons on Critical Thinking. One respondent said that debate and critical reflections will be added to class. One respondent reported that she is constantly innovating new approaches, including “an array of assignments designed to make literature and writing applicable to the students' experience outside of the classroom and include blogs, proposals to the community, investigative reporting, interview-based essays, etc.” One instructor has started asking students to think about what the next question would be, and using a problem/solution methodology for framing each segment of the course. The First Year Seminars will be incorporating more critical thinking instruction and assignments: “Our overarching approach will be to provide students with the language of critical thinking, (e.g., application, analysis, synthesis, multidimensional thinking) and to revisit this language regularly throughout the semester” by developing end-of-chapter assignments and weekly journal prompts. One respondent who is not a faculty member reported that he “interviewed faculty to identify how critical thinking was developed in coursework and class activities for inclusion in stories, features, etc. in university publications.”

Respondents provided a number of very specific changes with regard to Information Literacy: incorporation of specific tasks leading up to a major class project; incorporating explicit connections to IL in a research assignment; the addition of an annotated bibliography to an oral presentation assignment, adding "[t]hey do this in several stages so they can learn how to differentiate between scholarly and popular sources, and how to locate and identify quality online and in-library sources;" the addition of instruction on discerning the credibility and appropriateness of websites (rather than not allowing websites as sources); the addition of library sessions; and the incorporation of unique ways to obtain information (blogs, speakers, events and fairs). Other more general responses included more guidance regarding sources and making sure that the course outcomes are in line with the University Studies learning outcomes.

The main change to written communication was the incorporation of the writing process into the course, with emphasis on feedback (3 out of the 5 responses). One response noted that CHM 101 would now include a complete laboratory report (previously, students were required to write laboratory experiment responses which were not in the form of a complete report). One respondent also mentioned that the courses would provide more writing opportunities, and another listed better rubrics. The respondent discussed under Information Literacy who added an annotated bibliography also mentioned under written communication that "[t]hey do this in several stages so they can benefit from written feedback and continue to improve their writing and citation skills."
Overall, this is a substantial list of changes to course curriculum that should improve student abilities in critical thinking, information literacy, and written communication.

New Recommendations

Based on the assessment findings from 2011-2012 and the results of the review of actions towards previous recommendations, the following recommendations were adopted by the Learning Assessment Council on October 2, 2012:

- The General Education Assessment office will disaggregate the Inquiry data for dimension 6 to analyze possible differences between courses and/or sections.
- The LAC will distribute a "Did You Know" email to faculty with the results from this and other Inquiry studies and ask the faculty to share examples of what they are doing/might do regarding teaching the significance of limitations and implications to inquiry (IN6).
- The General Education Assessment office will work with the Department of Foreign Languages and Literatures to devise common rubrics for University Studies foreign language courses.
- The General Education Assessment office will provide individual results to each instructor that participated in the Diversity and Global Citizenship samples, along with scorer comments.
- A Director of University Studies position should be created and filled by July 1, 2013.

REFERENCES AND RESOURCES

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. (2nd edition). Thousand Oaks, CA: Sage Publications.

Rhodes, T. ed. (2010). Assessing Outcomes and Improving Achievement Tips and Tools for Using Rubrics. Washington, DC: Association of American Colleges and Universities. (Copies of the VALUE Rubrics can be found at http://www.aacu.org/value/rubrics/index_p.cfm?CFID=33263360&CFTOKEN=78277616)

Siefert, L. (2010) *General Education Assessment Spring 2010 Report*. University of North Carolina Wilmington internal document. http://www.uncw.edu/assessment/Documents/GeneralEducationAssessmentReportSp2010Final.pdf

University of North Carolina Wilmington. (2011). *UNCW Learning Goals*. adopted by Faculty Senate March 17, 2009 and modified on January 18, 2011. http://www.uncw.edu/assessment/uncwLearningGoals.html

University of North Carolina Wilmington. (2009). *Report of the General Education Assessment Committee*, March 2009. http://www.uncw.edu/assessment/Documents/General%20Education/GenEdAssessmentCommitteeReportMarch2009.pdf

APPENDIX A UNIVERSITY STUDIES CURRICULUM MAP

University Studies Component Student Learning Outcomes and UNCW Learning Goals

		Creative Inquiry		Critical Thinking		Thoughtful Expression		Responsible Citizenship	
		Foundational Knowledge	**Inquiry**	**Information Literacy**	**Critical Thinking**	**Thoughtful Expression**	**Second Language**	**Diversity**	**Global Citizenship**
Foundations	**Composition**		CMP2, CMP3	CMP3	CMP1, CMP2, CMP3, CMP4	CMP1, CMP2, CMP3, CMP4			
	Fresh Seminar		FS2	FS1	FS3	FS4			
	Foreign Language	SL1, SL2, SL4	SL4		SL1, SL2, SL3, SL4		FL1, FL2, FL3	SL4	SL3, SL4
	Lifespan Wellness	W1, W2, W3, W4			W1				
	Mathematics and Statistics	MS1, MS2	MS1, MS2	MS2	MS1, MS2, MS3	MS3			
Approaches and Perspectives	**Aesthetic, Interpretive, and Literary Perspectives**	AIL1	AIL1	AIL1	AIL1, AIL2, AIL3	AIL1		AIL2, AIL3	
	Historical and Philosophical Approaches	HPA1	HPA1, HPA3, HPA4	HPA2	HPA2, HPA4			HPA3	HPA4
	Living in a Global Society	GS1, GS2	GS2		GS2			GS2	GS2, GS3
	Living in Our Diverse Nation	LDN1, LDN3	LDN3	LDN2, LDN4	LDN2, LDN4			LDN1, LDN3, LDN4	
	Scientific Approaches to the Natural World	SAN1, SAN2	SAN1, SAN2	SAN2	SAN1, SAN2, SAN3	SAN3			
	Understanding Human Institutions and Behaviors	HIB1		HIB2	HIB2, HIB3, HIB4				HIB4
Common Requirements	**Information Literacy**		IL1, IL3	IL1, IL2, IL3, IL4, IL5	IL1, IL2, IF3, IL4, IL5	IL4			
	Quantitative Logical Reasoning	QRE1, QRE2	QRE1, QRE2 LOG1, LOG2, LOG3	QRE1, QRE2	QRE1, QRE2, QRE3 LOG1, LOG2, LOG3	QRE3 LOG3			
	Writing Intensive	WI1, WI5	WI3	WI2, WI3, WI5	WI2, WI4, WI5	WI3, WI4, WI5			
	Capstone								

Shaded items are the focus of General Education Assessment activities during present cycle (Fall 2011 to Spring 2014).

APPENDIX B A NOTE ON INTERRATER RELIABILITY MEASURES

There is much debate about the best means of measuring interrater reliability. There are many measures that are used. Some differences in the measures are due to the types of data (nominal, ordinal, or interval data). Other differences have to do with what is actually being measured. Correlation coefficients describe *consistency* between scorers. For example, if Scorer 1 always scored work products one level higher than Scorer 2, there would be perfect correlation between them. You could always predict one scorer's score by knowing the other's score. It does not, however, yield any information about *agreement.* A value of 0 for a correlation coefficient indicates no association between the scores, and a value of 1 indicates complete association. Spearman rho rank order correlation coefficient is an appropriate correlation coefficient for ordinal data.

Percent agreement measures exactly that—the percentage of scores that are exactly the same. It does not, however, account for chance agreement. Percent adjacent measures the number of times the scores were exactly the same plus the number of times the scores were only one level different. Percent adjacent lets the researcher know how often there is major *disagreement* between the scorers on the quality of the artifact.

Krippendorff's alpha is a measure of agreement that accounts for chance agreement. It can be used with ordinal data, small samples, and with scoring practices where there are multiple scorers. A value of 0 for alpha indicates only chance agreement, and a value of 1 indicates reliable agreement not based on chance. Negative values indicate "systematic disagreement" (Krippendorff, 2004).

Determining acceptable values for interrater reliability measures is not easy. Acceptable levels will depend on the purposes that the results will be used for. These levels must also be chosen in relationship to the type of scoring tool or rubric, and the measure of reliability being used. In this case, the tool is a "metarubric," a rubric that is designed to be applied across a broad range of artifacts and contexts. This type of instrument requires more scorer interpretation than rubrics designed for specific assignments. For consistency measures, such as correlation coefficients, in a seminal work, Nunnally states that .7 may suffice for some purposes whereas for other purposes "it is frightening to think that any measurement error is permitted" (Nunnally, 1978, pp.245-246). The standard set for Krippendorff's alpha by Krippendorff himself is .8 to ensure that the data are at least similarly interpretable by researchers. However, "where only tentative conclusions are acceptable, alpha greater than or equal to .667 may suffice" (Krippendorff, 2004, p. 241). In the present context, we should aim for values of at least .67, with the recognition that this could be difficult given the broad range of artifacts scored with the metarubrics.