GENERAL EDUCATION ASSESSMENT

Report for Academic Year 2012-2013

Prepared by

Dr. Linda Siefert
Director of General Education Assessment

Lea Bullard
Assistant Director of General Education Assessment

October 2013

This page purposely blank.

Acknowledgements

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

Judy Kinney from the Office of Institutional Research and Assessment, who provided demographich and preparedness data.

Caitlin Northcutt, graduate assistant for General Education Assessment

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 2012 – 2013. This was the second 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 16 courses within five University Studies components.

FINDINGS

FOUNDATIONAL KNOWLEDGE

In Fall 2012, 377 student final exams in PED 101 were sampled from four different course delivery types and the responses to 13 questions on those exams were analyzed. For all students, the average total score on the 13 questions sampled was 84.2%. The percentage of correct student answers was above 75% for all but two questions. There were significant differences (p= .05) between course delivery types for four questions. For each of these four questions, students in the Online Lecture, Face-to-Face Lab delivery method scored higher than the other three delivery types.

INFORMATION LITERACY

In Fall 2012 and Spring 2013, 69 student work products from PSY 105 and SED 372were scored on the UNCW Information Literacy rubric. The percent of students scored at or above the proficiency level were: IL1 Determine Extent of Information Needed – 84.0%; IL2 Access Needed Information – 82.6%; IL3 Evaluate Information and Sources – 73.9%; IL4 Use Information Effectively – 88.4%; and IL5 Access and Use Information Ethically – 82.6%. IL3 Evaluate Information and Sources has been the lower-scoring dimension in past studies as well.

CRITICAL THINKING

In Fall 2012 and Spring 2013, 175 student work products from ANTL 207, COM 160, ECN 222, ENG 230, FST 110, PSY 105, and THR 121 were scored on a rubric based on the VALUE Critical with some minor modifications made based on past scorer feedback. The percent of students scored at or above the proficiency level were: CT1 Explanation of Issues – 57.2%; CT2a Evidence: Selecting and Using – 53.8%; CT2b Evidence: Critically Examining for Viewpoint – 34.8%; CT3a Influence of Assumptions – 32.0%; CT3b Influence of Context – 35.4%; CT4 Student's Position 39.0%; and CT5 Conclusions and Related Outcomes – 35.4%. Scorers all agreed that CT1 and CT2a fit the assignments while the other—and also the lower-scoring dimensions—were not a good fit with the assignments scored.

Thoughtful Expression (Written)

In Spring 2013, 187 student work products from six courses were scored on the AAC&U VALUE rubric. The percent of students scored at or above the proficiency level were: WC1 Context and Purpose of Writing – 81.3%; WC2 Content Development – 80.2%; WC3 Genre and Disciplinary Conventions – 79.1%; WC4 Sources and Evidence – 78.5%; and WC5 Control of Syntax 86.5%. The classes from which the student work was selected included 100-level (56.6%), 300-level (16.6%), and 400-level (26.8%) courses., though there was no meaningful difference in the scores between the 100-, 300-, or 400-level courses, or related to the number of credit hours completed by students.

Second Language

In Fall 2012, 106 student oral exams from on a locally-created Second Language Speaking and Listening Rubric; 25 from FRH 201 and 81 from SPN 102 and 201. The percentages of students scored at or above the proficiency level for were: FRH Oral 1 Listening Comprehension—96.0%; FRH Oral 2 Pronunciation—93.0%; FRH Oral 3 Vocabulary—56.0%; FRH Oral 4 Grammar—48.0%; FRH Oral 5 Fluency 76.0%; SPN Oral 1 Listening Comprehension—88.9%; SPN Oral 2 Pronunciation—95.1%; SPN Oral 3 Vocabulary, Variety of Items and Expressions—49.4%; SPN 4 Oral Vocabulary Proper Use—90.1%; SPN 5 Oral Grammar—65.4%; SPN 6 Oral Fluency—75.3%. Listening Comprehension and Pronunciation were scored highly for both language, and the division of the Vocabulary dimension into variety and proper use seems to had important information.

Recommendation

Based on the discussion conclusions that (1) more instructions were needed in assignments to elicit higher performance in WC2 Content Development, WC3 Genre and Disciplinary Conventions, and WC4 Sources and Evidence and (2) assignments including directions such as “critically analyze,” “include rationale,” “summarize,” “evaluate,” “critique,” and “elaborate” produced higher-scoring work, the following recommendation was adopted by the University Studies Advisory Board:

The results and analysis of the assessment process must be disseminated more purposefully and broadly so that faculty members can address these findings in their courses. A team consisting of the Associate Vice Chancellor and Dean of Undergraduate Studies, the Director of General Education Assessment, the Chair of the University Studies Advisory Committee, and the Undergraduate Studies Liaison to University Studies will present findings and suggestions at each department/school faculty meeting during the 2014-2015 academic year. Cumulative results from 2011-2013 for Written Communication were chosen for the first round of presentations, as they are relevant to all courses.

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 and the University Studies Advisory Board provide 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 2012 – 2013. In Fall 2012 and Spring 2013, the following learning goals were assessed: Foundational Knowledge, Information Literacy, Critical Thinking, Thoughtful Expression (Written), and Second Language. 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 and new recommendations.

Methodology

For the purposes of this report, general education assessment activities in academic year 2012 – 2013 are divided into five areas: assessment of student learning in Foundational Knowledge, Information Literacy, Critical Thinking, Thoughtful Expression (Written), and Second Language.

The following questions were examined:

- What are the overall abilities of students taking University Studies courses with regard to the UNCW Learning Goals of Foundational Knowledge, Information Literacy, Critical Thinking, Thoughtful Expression (Written), and Second Language?
- What are the relative strengths and weaknesses within the subskills of those goals?
- Are there any differences in performance based on course delivery method or 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. One 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 (for all learning goals except Foundational Knowledge). 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 Information Literacy, Critical Thinking, and Thoughtful Expression (Written), 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. The versions of each of the rubrics that were used in the study are located in the appendices of each chapter.

A multiple-choice assessment was used for assessing Foundational Knowledge.

Benchmarks

The VALUE rubrics and most locally created rubrics are designed on a 0 to 4 scale. According to AAC&U, "the capstone [4] level reflects the demonstration of achievement for the specific criterion for a student who graduates with a baccalaureate degree. Milestones [2 and 3] suggest key characteristics of progressive learning as students move from early in their college experience to the completion of the baccalaureate degree" (Rhodes, 2010, p.2). Based on the design of these rubrics, UNCW uses the capstone level 4 as the benchmark for attainment of graduating seniors. For first- and second-year students assessed in lower-level general education courses, the milestone level 2 is the benchmark for achievement. The rationale for this is that performance at the milestone level 2 indicates that they are on track for achieving a level 4 by the time of graduation. Most locally-created rubrics were designed to follow these same levels. However, for the French and Spanish Listening and Speaking Rubric the benchmark for general education competency is 2.

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 2012-2013 studies, discipline-specific scorers were solicited for Second Language, whereas scorers were solicited from all departments for Information Literacy, Critical Thinking, and Thoughtful Expression (Written). 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. 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 work products independently and not in a scoring session). Scorers were assigned to groups. Scoring of each assignment began with the group 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. Learning-goal experts were available in each scoring event to answer any questions the faculty scorers had about scoring the student work product against the metarubric.

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 Lifetime Wellness component.

SUMMARY OF FINDINGS

In Fall 2012, the lead faculty teaching PED 101 Physical Activity and Wellness selected 13 questions on the final exam that aligned with the student learning outcome for the course related to identifying the thoughts, attitudes, choices, and behaviors associated with lifelong health and wellness. All questions were either multiple-choice or True/False-type questions. Therefore, the student responses were scored as either correct or incorrect. Students taking the final in four different course delivery types were included, totaling 377 students. Table 2.1 provides the results.

Table 2.1 Foundational Knowledge Lifetime Wellness Results

Question	% of Students answering correctly				
	Online lecture, F to F lab (N=114; 4 sections)	F to F lecture, F to F lab (N=58, 2 sections)	Online lecture, Online lab (N=117, 4 sections)	Online lecture, Web enhanced lab (N=88, 3 sections)	**All Students**
1	51%	28%	46%	38%	**42.7%**
2	99%	100%	100%	100%	**99.7%**
3	100%	100%	100%	100%	**100%**
4	97%	97%	93%	92%	**94.4%**
5	97%	98%	99%	99%	**98.3%**
6	98%	100%	99%	97%	**98.4%**
7	94%	86%	88%	80%	**87.5%**
8	86%	83%	90%	84%	**86.1%**
9	50%	45%	43%	34%	**43.2%**
10	100%	97%	96%	100%	**98.2%**
11	84%	78%	81%	73%	**79.6%**
12	86%	66%	76%	70%	**76.1%**
13	93%	86%	88%	93%	**90.4%**
Average Total Score	**87.3%**	**81.7%**	**84.6%**	**81.4%**	**84.2%**

For all students, the average total score on the 13 questions sampled was 84.2%. The percentage of correct student answers was above 75% for all but two questions.

Comparisons between Delivery Mode

Four different modes of instruction were used to deliver PED 101: online lecture, face-to-face lab; face-to-face lecture and lab; online lecture and lab; and online lecture, web-enhanced lab. In this sample, most students were in either online lecture and face-to-face labs (n=114) or online lecture and labs (n=117). There were 88 students in the sample in online lecture and web-enhanced labs and 58 students in face-to-face lecture and labs.

There was no significant difference in the score distributions between course delivery type for nine of the 13 questions. There were significant differences (p= .05) between delivery types for questions 1 (number of cardio mins/week for health), 7 (intrinsic motivation to exercise), 10 (product of healthy behaviors), and 12 (priority lifestyles). For each of these four questions, students in the Online Lecture, Face-to-Face Lab delivery method scored higher than the other three delivery types.

Discussion

In general, the percentage of students answering correctly on each of the 13 questions was high. The exceptions to this statement are questions 1 (number of cardio mins/week for health) and 9 (process-oriented goal), for which 51% or fewer students answered correctly across any delivery type. The lack of statistically significant differences across most items and on the whole test suggests that the differences between course delivery type are fairly minor, although the online lecture, face-to-face- lab students performed better on a few questions.

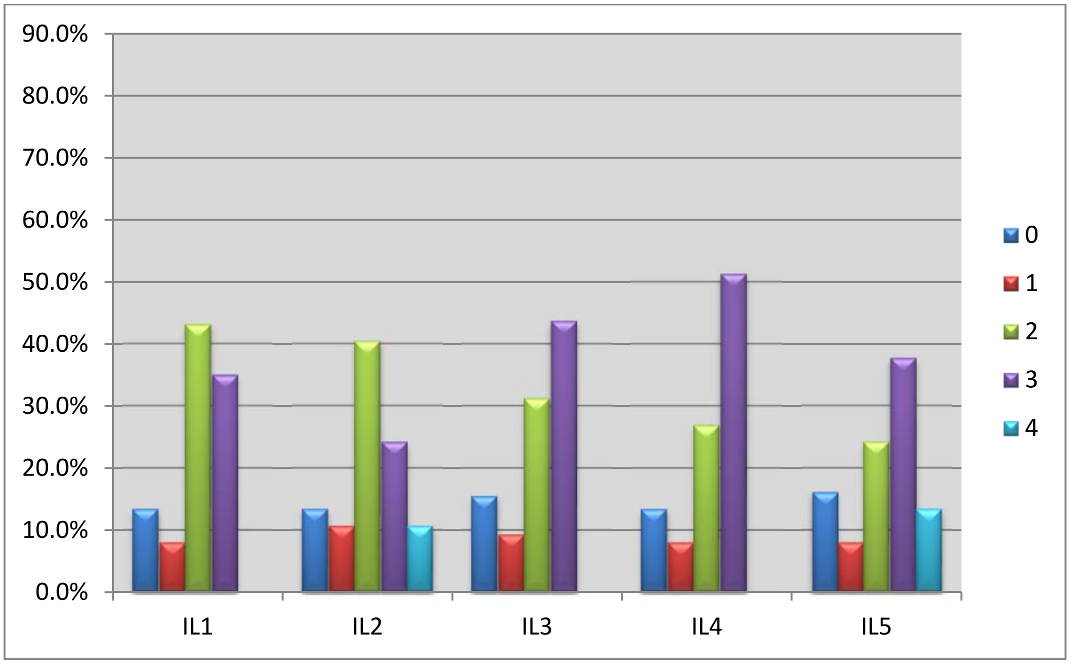
3. INFORMATION LITERACY

The UNCW Information Literacy learning goal is for students to locate, evaluate, and effectively use information by applying a variety of academic and technological skills. For purposes of this learning goal, information literacy is characterized by the ability to determine the extent of the information needed, accessing the needed information, critically evaluating the information, organizing the information to accomplish a specific purpose, and using the information ethically and legally (UNCW Learning Goals, 2011). The Information Literacy rubric is based largely on the AAC&U VALUE rubric and details five dimensions of knowledge related to information literacy. The Information Literacy rubric can be found in Appendix 3.A at the end of this chapter. In this study, work was sampled from the Understanding Human Institutions and Behaviors and Writing Intensive components of University Studies.

SUMMARY OF SCORES BY DIMENSION

Four faculty scorers scored 69 work products from three assignments across two courses from the Fall 2012 and Spring 2013 semesters: PSY 105 (two sections) and SED 372 (two sections). The assignments were all completed out-of-class and consisted of research papers, critique papers, and projects, some completed as groups and some individually. Eighteen work products (26.1%) were scored by multiple scorers. Figures 3.1.a and 3.1.b provide the score distributions for each dimension for work products that were scored on that dimension (i.e., work products with blank scores are not included).

INFORMATION LITERACY RESULTS BY DIMENSION FOR APPLICABLE SCORES ONLY, LOWER-DIVISION COURSES



0	13.5%	13.5%	15.6%	13.5%	16.2%
1	8.1%	10.8%	9.4%	8.1%	8.1%
2	43.2%	40.5%	31.3%	27.0%	24.3%
3	35.1%	24.3%	43.8%	51.4%	37.8%
4	0.0%	10.8%	0.0%	0.0%	13.5%
25th %tile	2	1.5	1.25	2	1.5
50th %tile	2	2	2	3	3
75th %tile	3	3	3	3	3
Mode	2	2	3	3	3
N	37	37	32	37	37

Figure 3.1.a Distribution of Scores for Information Literacy in Lower-Division Courses, Applicable Scores Only

RESULTS BY DIMENSION FOR LOWER-DIVISION COURSES

IL1 Determine Extent of Information Needed

Just fewer than one in seven student work products were scored a zero on this dimension. One in 12 work products indicated difficulty defining the scope of the research question or thesis (scores of one). Slightly fewer than half of the work products incompletely defined the scope of the research question or thesis but did determine some key concepts (scores of two). Just over one-third of the work products defined the scope of the research question or thesis completely and determined key concepts (scores of three). No student work products from lower-division courses effectively defined the scope of the research question and effectively determined key concepts (scores of four).

IL2 Access Needed Information

One in seven student work products received a score of zero for this dimension. One in 10 work products accessed information randomly or retrieved information that lacked relevance and quality (scores of one). Four in 10 student work products accessed information using simple search strategies and retrieved information from some relevant, though limited and similar sources (scores of two). Just under one-fourth of work products accessed information using a variety of search strategies and from relevant sources, while also demonstrating the ability to refine the search strategy (scores of three). One in 10 work products accessed information using effective, well-designed search strategies and from the most appropriate information sources (scores of 4).

IL3 Evaluate Information and Sources

Just fewer than one in six student work products failed to show any evidence of evaluating information and it sources critically (scores of zero). One in 11 work products evidenced that information was taken from sources without any interpretation or evaluation of the materials and that the viewpoints of the authors were not subject to questioning (scores of one). Just less than one-third of the work products indicated that information was taken from source(s) with some interpretation/evaluation, but did not include a coherent analysis of the material and viewpoints of the authors were taken mostly as fact (scores of two). A little under one in two of work products demonstrated that information was taken from source(s) with enough interpretation or evaluation to develop a coherent analysis of the material, with viewpoints of authors questioned (scores of three). No student work products received scores of four for this dimension.

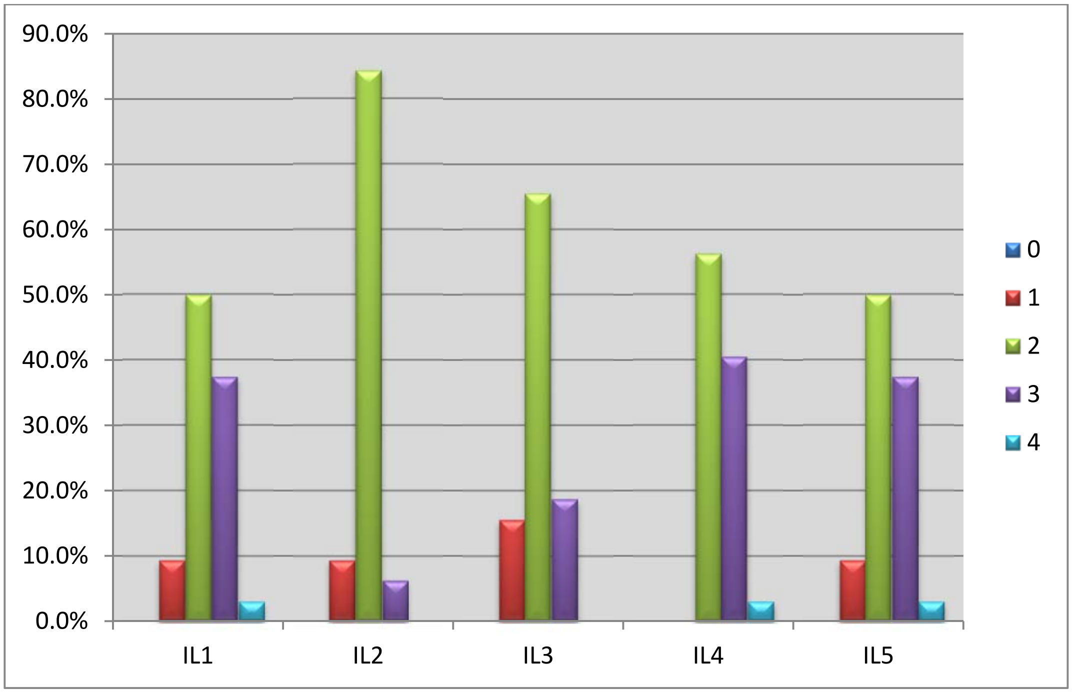
IL4 Use Information Effectively

Just over one in eight student work products received a score of zero for this dimension. One in 12 papers communicated and organized information from sources, but that information was not fully synthesized (scores of one). Just over one-quarter of the papers communicated, organized, and synthesized information from sources, achieving the intended purpose (scores of two). A little more than half of the papers received scores of three, indicating that the student communicated, organized, and synthesized information from sources and that the intended purpose was achieved. No papers communicated, organized, and synthesized information from sources, achieving the intended purpose with clarity and depth (score of four).

IL5 Access and Use Information Ethically

This dimension deals with ethical use of information. One in six work products received a score of zero. For any score greater than zero for this dimension, the differences in score level are based on the number of information-use strategies employed. Any score above zero also indicates that a work product demonstrates a full understanding of the ethical and legal restrictions on the use of published, confidential, and/or proprietary information. One in 12 work products received scores of one, indicating the work consistently used only one information use strategy (citations and references; choice of paraphrasing, summary, or quoting; using information in ways that are true to original context; and distinguishing between common knowledge and ideas requiring attribution). Just under one-fourth of the work products demonstrated two types of information-use strategies (score of two). A little over three in 10 work products showed evidence of three types of information-use strategies (score of three). Just less than one in seven work products showed evidence of all types of information-use strategies (score of four).

INFORMATION LITERACY RESULTS BY DIMENSION FOR APPLICABLE SCORES ONLY, UPPER-DIVISION COURSES



0	0.0%	0.0%	0.0%	0.0%	0.0%
1	9.4%	9.4%	15.6%	0.0%	9.4%
2	50.0%	84.4%	65.6%	56.3%	50.0%
3	37.5%	6.3%	18.8%	40.6%	37.5%
4	3.1%	0.0%	0.0%	3.1%	3.1%
25th %tile	2	2	2	2	2
50th %tile	2	2	2	2	2
75th %tile	3	2	2	3	3
Mode	2	2	2	2	2
N	32	32	32	32	32

Figure 3.1.b Distribution of Scores for Information Literacy in Upper-Division Courses, Applicable Scores Only

RESULTS BY DIMENSION FOR UPPER-DIVISION COURSES

IL1 Determine Extent of Information Needed

No work products were scored a zero on this dimension. One in 11 student work products indicated difficulty defining the scope of the research question or thesis (scores of one). One-half of the work products incompletely defined the scope of the research question or thesis but did determine some key concepts (scores of two). Slightly over one-third of the work products defined the scope of the research question or thesis completely and determined key concepts (scores of three). Only one out of 30 work product effectively defined the scope of the research question and effectively determined key concepts (scores of four).

IL2 Access Needed Information
No student work products received a score of zero for this dimension. One in 11 work products accessed information randomly or retrieved information that lacked relevance and quality (scores of one). Just over five in six student work products accessed information using simple search strategies and retrieved information from some relevant, though limited and similar sources (scores of two). One in 16 work products accessed information using a variety of search strategies and from relevant sources, while also demonstrating the ability to refine the search strategy (scores of three). No student work products accessed information using effective, well-designed search strategies and from the most appropriate information sources (scores of 4).

IL3 Evaluate Information and Sources
No work products failed to show any evidence of evaluating information and it sources critically (scores of zero). Slightly over one in seven work products evidenced that information was taken from sources without any interpretation or evaluation of the materials and that the viewpoints of the authors were not subject to questioning (scores of one). Just under two-thirds of the work products indicated that information was taken from source(s) with some interpretation/evaluation, but did not include a coherent analysis of the material and viewpoints of the authors were taken mostly as fact (scores of two). Slightly under one in five work products demonstrated that information was taken from source(s) with enough interpretation or evaluation to develop a coherent analysis of the material, with viewpoints of authors questioned (scores of three). No student work products received scores of four for this dimension.

IL4 Use Information Effectively
No student work products received a score of zero for IL4. Likewise, no work showed communicated and organized, though incompletely synthesized, information from sources (scores of one). Just over half of the papers communicated, organized, and synthesized information from sources, achieving the intended purpose (scores of two). Just over one-third of the papers received scores of three, indicating that the student communicated, organized, and synthesized information from sources and that the intended purpose was achieved. One paper out of 30 communicated, organized, and synthesized information from sources, achieving the intended purpose with clarity and depth (score of four).

IL5 Access and Use Information Ethically
This dimension deals with ethical use of information. No student work products were scored a zero for this dimension. For any score greater than zero for this dimension, the differences in score level are based on the number of information-use strategies employed. Any score above zero also indicates that a work product demonstrates a full understanding of the ethical and legal

restrictions on the use of published, confidential, and/or proprietary information. One in 11 work products received scores of one, indicating the work consistently used only one information use strategy (citations and references; choice of paraphrasing, summary, or quoting; using information in ways that are true to original context; and distinguishing between common knowledge and ideas requiring attribution). Half of the work products demonstrated two types of information-use strategies (score of two). Slightly over one-third of the work products showed evidence of three types of information-use strategies (score of three). One in 30 work products showed evidence of all types of information-use strategies (score of four).

Correlation between Dimensions

All dimensions were correlated at the .001 level of significance. See Appendix 3.B for the complete presentation of dimension correlation coefficients.

Demographic and Preparedness Findings

There were no statistically significant difference between the means, medians, and the score distributions of the different race/ethnicity categories, male and female students, transfer vs. UNCW-start students, and honors vs. 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 (45.2% of the sample), 31 – 60 credit hours (3.2% of the sample), 61 – 90 (21.0% of the sample), and over 90 credit hours (30.6% 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 for any of the Information Literacy dimensions. Looking at Spearman rho correlation coefficients, the number of total hours completed was positively correlated with IL1 Determine Extent of Information Needed (.346**) and IL4 Use Information Effectively (.272*).

There were some significant correlations with GPA and SAT-Verbal scores and the Information Literacy dimensions. GPA was positively correlated with IL2 Access Needed Information (.460**) and with IL5 Access and Use Information Ethically (.413*). IL5 was also positively correlated with SAT-Verbal (.385*). There were no significant correlations between SAT-Math and ACT scores and the Information Literacy dimension scores.

Comparisons between upper- and lower-division, courses, and assignment types

Only one significant difference between upper- and lower division courses was seen. Parametric tests showed statistically significant differences between the levels for the ranges of IL1 and IL4, with the range for both dimensions 0-3 for lower-division course and 1-4 for the upper-division courses.

Comparisons between University Studies components

Student work that was scored for Information Literacy was sampled from courses in the Understanding Human Institutions and Behaviors and Writing Intensive components of University Studies. Since these align exactly with lower-division and upper-division courses, see the section above.

Interrater Reliability

For each group of scorers, there were a number of common papers scored so that interrater reliability could be assessed (a total of 18 or 26.2%). Table 3.1 shows the reliability measures for Information Literacy.

Table 3.1 Interrater Reliability for Information Literacy

Dimension	N	Percent Agreement	Plus Percent Adjacent	Krippendorff's Alpha
IL1 Determine Extent of Information Needed	18	55.6%	94.4%	.521
IL2 Access Needed Information	18	55.6%	100%	.771
IL3 Evaluate Information and Sources	15	53.3%	93.3%	.413
IL4 Use Information Effectively	18	66.7%	100%	.722
IL5 Access and Use Information Ethically	18	55.6%	94.4%	.674

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 Krippendorff's Alpha measure scorer agreement. The UNCW benchmark is .67 for Krippendorff's Alpha. See Appendix B of this 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, there are three dimensions of the rubric that meets these standards, IL2, IL4, and IL5. Looking at percent agreement plus adjacent (that is, the scores that were within one level of each other), we find that all dimensions had greater than 90% of scores in agreement or within one level of each other.

Scorer Feedback on Rubric and Required Assumptions

Scorer opinion about the fit of the rubric dimensions to the assignment was mixed. For no dimension did all scorers agree that it fit equally with the assignment; there was generally a split in scorer opinion between “fit well” and “fit with guidelines”. For two dimensions, IL3 Evaluate Information and its Sources Critically and IL5 Access and Use Information Ethically and Legally, scorer opinion was mixed between “fit well”, “fit with guidelines”, and “did not fit”.

In response to the open-ended questions about the rubric dimensions that did not fit the assignment, all issues centered on the assignment itself, some being group projects, and evaluating the research sources. Comments on specific assignments included “the assignment was not a good fit for information literacy” and “assignment is just not detailed enough to make good judgments”. IL3 Evaluate Information and its Sources Critically seemed to be a problem because multiple scorers claimed it did not work for this type of assignment or they could not critically assess the sources.

When asked if there was a specific quality criterion that was problematic, one scorer commented that it was hard to determine what IL2 Access the Needed Information and IL3 Evaluate Information and its Sources Critically asked for and how scorers could accurately assess these dimensions. Two other scorers commented that the rubric fit “too well” and because it was a group assignment made it difficult to assess the individual student abilities. Only one scorer offered a suggestion for improvements that could be made to the rubric. They requested that the rubric itself be part of the assignment.

The last section of open-ended questions involved the specific language used in the instruction and how it affected student responses. The first question was on overall extent and specificity of instructions and how they affected student responses. The majority of scorers felt that the instructions were very complete and well written. They felt that student responses were directed clearly from the instructions. One scorer felt that improvement was needed in defining “scientific sources”. The last open-ended question concerned the parts of the assignment that were particularly the most effective in eliciting evidence of student information literacy. Scorers felt that statements such as, “synthesize literature and research, analyze ideas, and reflect,” and “include the following components” helped guide student responses. Verbs such as, “compile, summarize, and elaborate” gave students direction but one scorer did point out that phrases such as “critical analysis” are helpful, but only if the students have understanding of what it means.

Discussion

Table 3.2 shows the percent of work products scored at or above the benchmark levels.

Table 3.2 Information Literacy Percent of Sample Scored at or above 2 and 3

Lower-Division Courses		
Dimension	% of Work Products Scored Two or Higher	% of Work Products Scored Three or Higher
IL1 Determine Extent of Information Needed	78.3%	35.1%
IL2 Access Needed Information	75.6%	35.1%
IL3 Evaluate Information and Sources	75.1%	43.8%
IL4 Use Information Effectively	78.4%	51.4%
IL5 Access and Use Information Ethically	75.6%	51.3%
Upper-Division Courses		
Dimension	% of Work Products Scored Two or Higher	% of Work Products Scored Three or Higher
IL1 Determine Extent of Information Needed	90.6%	40.6%
IL2 Access Needed Information	90.7%	6.3%
IL3 Evaluate Information and Sources	84.4%	18.8%
IL4 Use Information Effectively	100.0%	43.7%
IL5 Access and Use Information Ethically	90.6%	40.6%

For the lower-division courses, the benchmark achievement is defined as rubric score level two. A majority of work products from the lower-division courses were scored at this benchmark level or better. For the upper-division courses, while there is no benchmark, a score of three suggests that the students in a 300-level course are poised to be able to reach the graduating senior benchmark of four with additional opportunities to practice. For no dimension did a majority of work products score at the level three for upper-division courses. Dimensions IL1, IL4, and IL5 showed the highest percentage of work products scoring a three or above, around 40%. IL3 had a much lower percentage of work products at these levels, with just under 20% of papers scoring a level three or better. IL2 was the dimension with the most need for improvement, with fewer than 10% of work products scoring a three or better.

There were no statistical differences found in the scores on any dimension between the two levels of courses (upper- vs. lower-division). Statistical differences between student total hours completed, or class level, were not analyzed because some of the work products were group projects on which many students worked. However, these results indicate that students in this study achieved at about the same levels on information literacy dimensions, regardless of course level.

Looking back to past studies of students’ information literacy skills, IL 3 Evaluate Information and Sources has been a historically low-scoring dimension for UNCW students. This was only the case for the upper-division scores in this study. Scorers’ feedback indicates that the assignments scored did not address these dimensions thoroughly, which may be a contributing factor to the lower scores.

APPENDIX 3.A INFORMATION LITERACY RUBRIC

INFORMATION LITERACY VALUE RUBRIC

for more information, please contact value@aacu.org

Association of American Colleges and Universities

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

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. - Adopted from the National Forum on Information Literacy

Framing Language

This rubric is recommended for use evaluating a collection of work, rather than a single work sample in order to fully gauge students' information skills. Ideally, a collection of work would contain a wide variety of different types of work and might include: research papers, editorials, speeches, grant proposals, marketing or business plans, PowerPoint presentations, posters, literature reviews, position papers, and argument critiques to name a few. In addition, a description of the assignments with the instructions that initiated the student work would be vital in providing the complete context for the work. Although a student's final work must stand on its own, evidence of a student's research and information gathering processes, such as a research journal/diary, could provide further demonstration of a student's information proficiency and for some criteria on this rubric would be required

Information Literacy Rubric

Based largely on the AAC&U VALUE Rubric; for more information, please contact value@aacu.org

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

	Benchmark	Milestones		Capstone	Score
	1	2	3	4	
1. Determine the Extent of Information Needed	Has difficulty defining the scope of the research question or thesis. Has difficulty determining key concepts. Types of information (sources) selected do not relate to concepts or answer research question.	Defines the scope of the research question or thesis incompletely (parts are missing, remains too broad or too narrow, etc.). Can determine key concepts. Types of information (sources) selected partially relate to concepts or answer research question.	Defines the scope of the research question or thesis completely. Can determine key concepts. Types of information (sources) selected relate to concepts or answer research question.	Effectively defines the scope of the research question or thesis. Effectively determines key concepts. Types of information (sources) selected directly relate to concepts or answer research question.	
2. Access the Needed Information	Accesses information randomly and retrieves information that lacks relevance and quality.	Accesses information using simple search strategies and retrieves information from some relevant, though limited and similar, sources.	Accesses information using variety of search strategies and from relevant information sources. Demonstrates ability to refine search.	Accesses information using effective, well-designed search strategies and from most appropriate information sources.	
3. Evaluate Information and its Sources Critically	Information is taken from source(s) without any interpretation/evaluation of the material; viewpoints of authors are taken as fact, without question.	Information is taken from source(s) with some interpretation/evaluation, but not a coherent analysis of the material; viewpoints of authors are taken as mostly fact, with little questioning.	Information is taken from source(s) with enough interpretation/evaluation to develop a coherent analysis of the material; viewpoints of authors are subject to questioning.	Information is taken from source(s) with enough interpretation/evaluation to develop a comprehensive analysis of the material; viewpoints of authors are questioned thoroughly.	
4. Use Information Effectively to Accomplish a Specific Purpose	Communicates information from sources. The information is fragmented and/or used inappropriately (misquoted, taken out of context, or incorrectly paraphrased, etc.), so the intended purpose is not achieved.	Communicates and organizes information from sources. The information is not yet synthesized, so the intended purpose is not fully achieved.	Communicates, organizes and synthesizes information from sources. Intended purpose is achieved.	Communicates, organizes and synthesizes information from sources to fully achieve a specific purpose, with clarity and depth	
5. Access and Use Information Ethically and Legally	Consistently uses one of the following information use strategies: • use of citations and references, • choice of paraphrasing, summary, or quoting, • using information in ways that are true to original context, • distinguishing between common knowledge and ideas requiring attribution; AND demonstrates a full understanding of the ethical and legal restrictions on the use of published, confidential, and/or proprietary information.	Consistently uses two of the following information use strategies: • use of citations and references, • choice of paraphrasing, summary, or quoting, • using information in ways that are true to original context, • distinguishing between common knowledge and ideas requiring attribution; AND demonstrates a full understanding of the ethical and legal restrictions on the use of published, confidential, and/or proprietary information.	Consistently uses three of the following information use strategies: • use of citations and references, • choice of paraphrasing, summary, or quoting, • using information in ways that are true to original context, • distinguishing between common knowledge and ideas requiring attribution; AND demonstrate sa full understanding of the ethical and legal restrictions on the use of published, confidential, and/or proprietary information.	Consistently uses all of the following information use strategies: • use of citations and references, • choice of paraphrasing, summary, or quoting, • using information in ways that are true to original context, • distinguishing between common knowledge and ideas requiring attribution; AND demonstrates a full understanding of the ethical and legal restrictions on the use of published, confidential, and/or proprietary information.	

NOTES:

APPENDIX 3.B CORRELATIONS BETWEEN INFORMATION LITERACY DIMENSIONS

Spearman rho Rank Order Correlation Coefficients

			IL1	IL2	IL3	IL4	IL5
Spearman's rho	IL1	Correlation Coefficient		.728**	.528**	.711**	.674**
		N		69	64	69	69
	IL2	Correlation Coefficient	.728**		.737**	.680**	.723**
		N	69		64	69	69
	IL3	Correlation Coefficient	.528**	.737**		.766**	.665**
		N	64	64		64	64
	IL4	Correlation Coefficient	.711**	.680**	.766**		.741**
		N	69	69	64		69
	IL5	Correlation Coefficient	.674**	.723**	.665**	.741**	
		N	69	69	64	69	

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

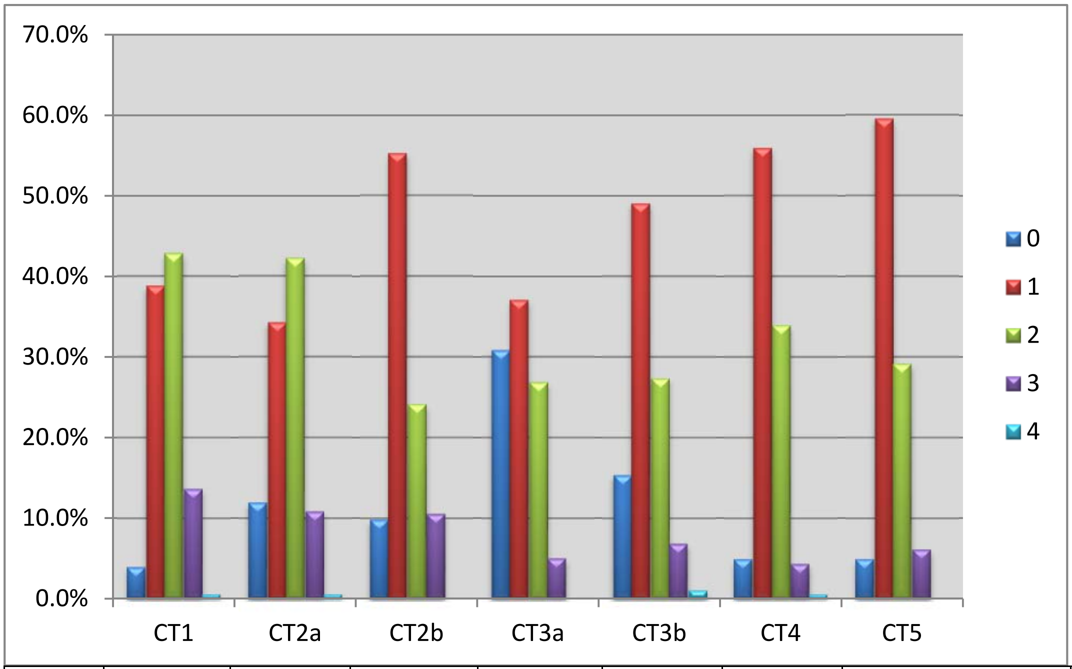
4. CRITICAL THINKING

The UNCW Critical Thinking learning goal is for students to use multiple methods and perspectives to critically examine complex problems. For purposes of this Learning Goal, "Critical thinking is 'skilled, active interpretation and evaluation of observations, communications, information and argumentation' (Fisher and Scriven, 1997). Critical thinking involves a clear explanation of relevant issues, skillful investigation of evidence, purposeful judgments about the influence of context or assumptions, reasoned creation of one's own perspective, and synthesis of evidence and implications from which conclusions are drawn" (UNCW Learning Goals, 2011). The rubric used to score this learning goal is based largely on the VALUE Critical Thinking rubric; based on feedback from scorers, some minor modifications were made to the rubric. This updated version of the Critical Thinking rubric can be found in Appendix 4.A at the end of this chapter. In this study, work was sampled from the Understanding Human Institutions and Behaviors and Aesthetic, Interpretive, and Literary Perspective components of University Studies.

SUMMARY OF SCORES BY DIMENSION

Seven faculty scorers scored 175 work products from nine assignments across seven courses from the Fall 2012 and Spring 2013 semesters: ANTL 207, COM 160, ECN 222, ENG 230, FST 110, PSY 105, and THR 121. ANTL 207, COM 160, ECN 222, and PSY 105 are courses approved for the Understanding Human Institutions and Behaviors component of University Studies. ENG 230, FST 110, and THR 121 are in the Aesthetic, Interpretive, and Literary Perspective component of University Studies. The types of assignments consisted of lab exercises, final exam essay questions, research papers, critique papers, and projects, some completed as groups and some individually. Sixty work products (34.3%) were scored by multiple scorers. Figure 4.1 provides the score distributions for each dimension for work products that were scored on that dimension (i.e., work products with blank scores are not included).

Critical Thinking Results by Dimension for Applicable Scores Only



	CT1	CT2a	CT2b	CT3a	CT3b	CT4	CT5
0	4.0%	12.0%	9.9%	30.9%	15.4%	5.0%	5.0%
1	38.9%	34.3%	55.3%	37.1%	49.1%	56.0%	59.6%
2	42.9%	42.3%	24.2%	26.9%	27.4%	34.0%	29.2%
3	13.7%	10.9%	10.6%	5.1%	6.9%	4.4%	6.2%
4	0.6%	0.6%	0.0%	0.0%	1.1%	0.6%	0.0%
25th %tile	1	1	1	0	1	1	1
50th %tile	2	2	1	1	1	1	1
75th %tile	2	2	2	2	2	2	2
Mode	2	2	1	1	1	1	1
N	175	175	161	175	175	159	161

Figure 4.1 Distribution of Scores for Critical Thinking, Applicable Scores Only

Results by Dimension

CT 1 Explanation of Issues

Scores were higher on this dimension than on the other Critical Thinking rubric dimensions. Fewer than one in 20 work products scored a zero on CT1. Over one-third of the work products received scores of one, indicating that the issue or problem to be considered critically was stated, but without clarification or description. Over four in 10 work products were scored at a level two for this dimension, indicating that the issue or problem to be considered was stated and described, but that there were some aspects left unexplored. Slightly more than one in eight work products stated, described, and clarified the issue or problem to the extent that understanding was not seriously impeded by omissions (scores of 3). Finally, one work product received a score of four on this dimension, indicating that the issue to be considered critically was clearly stated and comprehensively described.

CT2a Evidence: Selecting and Using

The scores on CT2a were the second-highest scores for Critical Thinking. About one in eight papers received scores of zero on this dimension. Just over one-third of the work products contained information taken from sources without any interpretation or evaluation (scores of one). Four in 10 of the work products provided information from sources along with some interpretation, but that evaluation was not enough to develop a coherent analysis or synthesis (scores of two). One out of 10 papers used information taken from sources with enough interpretation/evaluation to develop a coherent analysis or synthesis (scores of three). One paper evidenced information taken from sources with enough interpretation to develop a comprehensive analysis (score of four).

CT2b Evidence: Critically Examining for Viewpoint

One in 10 work products received scores of zero on this dimension. Over half of the work products received scores of one, indicating that the viewpoints of authors of the evidence cited were taken as fact, without question. One in four work products used evidence with little questioning of the viewpoints of the authors cited (scores of two). One in 10 papers indicated that the viewpoints of authors of the evidence used were subject to questioning (scores of three). No work products thoroughly questioned the viewpoints of authors of the evidence cited (scores of four).

CT3a Influence of Assumptions

The scores on this dimension were the lowest for Critical Thinking. Three out of every 10 work products scored a zero on this dimension. Almost four of every 10 work products showed an emerging awareness of either the student’s own assumptions or those of others (scores of one). Just over one in four students received scores of two on this dimension, meaning that the work questioned some assumptions and was perhaps more aware of others’ assumptions than of one’s own (or vice versa). One in twenty work products was scored at a level three, meaning that the student identified own and others’ assumptions when presenting a position. No work products were scored at a level four.

CT3b Influence of Context

The dimension had the second-lowest scores for Critical Thinking. Three in 20 papers received a zero for this dimension. One-half of the work products were scored at a level one for Influence of Context, meaning that those papers began to identify some contexts when presenting a position. Just over one-quarter of the papers received scores of two for this dimension, meaning that students identified several relevant contexts when presenting a position. Seven in one hundred papers identified several relevant contexts and discussed at least some aspects of their interconnectedness (scores of three). One percent of the papers were scored as a level four, meaning that the paper carefully evaluated the relevance of context when presenting a position.

CT4 Student's Position

Scores of zero were assigned to one in 20 papers for CT4 Student's Position. Over half of the papers scored stated a specific position, but that position was simplistic and obvious (scores of one). One-third of the work products received scores of two, indicating that the specific position that was stated acknowledged different sides of an issue. One in 20 papers stated a specific position that took into account the complexities of an issue, as well as acknowledging others' points of view within that position; these papers were assigned a score of three for CT4. Finally, one paper received a score of four for this dimension, meaning that the specific position stated was imaginative, taking into account the complexities of an issue, that the limits of that position were acknowledged, and that others' viewpoints were synthesized within the stated position.

CT5 Conclusions and Related Outcomes

One in 20 papers received scores of zero for this dimension. Three out of five papers were scored at a level one, meaning that the conclusion was inconsistently tied to some of the information discussed in the paper, and that related outcomes were oversimplified. For three out of 10 papers, the conclusion was logically tied to the information discussed though only information that fit the desired conclusion was introduced, and some related outcomes were identified (scores of two). Three in 50 papers scored a three on this dimension, meaning that the conclusion was logically tied to a range of information including opposing viewpoints, and that related outcomes were clearly identified. No scores of four were assigned for CT 5 Conclusions and Related Outcomes.

CORRELATION BETWEEN DIMENSIONS

All dimensions were correlated at the .001 level of significance. See Appendix 4.B for the complete presentation of dimension correlation coefficients.

DEMOGRAPHIC AND PREPAREDNESS FINDINGS

There were no statistically significant differences between the means, medians, and the score distributions of the different race/ethnicity categories, transfer vs. UNCW-start students, and honors vs. non-honors students. There were statistically significant differences between male and female students on CT1, CT2a, CT2b, CT3a, CT3b, and CT4, with females scoring higher on the six dimensions.

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 (24.7% of the sample), 31 – 60 credit hours (37.7% of the sample), 61 – 90 (22.8% of the sample), and over 90 credit hours (14.8% of the sample). Comparison of means (using ANOVA), medians (using Independent Samples test of medians) and distributions (using the Mann-Whitney U statistic) showed no statistically significant differences between the groups for any of the Critical Thinking dimensions. Looking at Spearman rho correlation coefficients, the number of total hours completed was not significantly correlated with any of the Critical Thinking dimensions. There were some significant correlations with GPA and SAT-Verbal scores and the Critical Thinking dimensions. GPA was positively correlated with all dimensions: CT1 (.343**), CT2a (.346**), CT2b (.207*), CT3A (.227*), CT3b (.277**), CT4 (.322**), and CT5 (.266**). CT2a and CT4 were also positively correlated with SAT-Verbal (.196* and .202*, respectively). ACT scores were positively correlated with CT2b (.517**). There were no significant correlations between SAT-Math and the Critical Thinking dimension scores.

Comparisons between courses and assignment types

The assignments, instructional setting, and instructor type of the work products collected were varied. Statistical analyses indicated differences between the scores of some of these groups. There was a statistical difference between the scores on CT2a, CT2b, CT3a, CT3b, and CT5 between in-class and out-of-class assignments, with the assignments that were completed out-of-class scoring higher on these dimensions. Assignments completed for classes taught by non-tenure-track faculty scored statistically higher on all dimensions. Assignments completed for online courses scored statistically higher on CT1, CT3a, CT3b, CT4, and CT5.

Comparisons between University Studies components

To assess critical thinking, work was sampled and scored from two components of University Studies: Understanding Human Institutions and Behaviors (n=103) and Aesthetic, Interpretive, and Literary Perspectives (n=72). There were statistical differences between the scores of work sampled from these two components. For all critical thinking dimensions, work sampled from Aesthetic, Interpretive, and Literary Perspectives was scored higher.

Interrater Reliability

For each group of scorers, there were a number of common papers scored so that interrater reliability could be assessed (60 or 34.3%). Table 4.1 shows the reliability measures for Information Literacy.

Table 4.1 Interrater Reliability for Critical Thinking

Dimension	N	Percent Agreement	Plus Percent Adjacent	Krippendorff's Alpha
CT 1 Explanation of Issues	58	34.5%	91.4%	.415
CT2a Evidence: Selecting and Using	58	32.8%	81.0%	.342
CT2b Evidence: Critically Examining for Viewpoint	53	30.2%	86.8%	.335
CT3a Influence of Assumptions	58	46.6%	86.2%	.524
CT3b Influence of Context	58	34.5%	89.7%	.500
CT4 Student's Position	54	59.3%	92.6%	.488
CT5 Conclusions and Related Outcomes	53	45.3%	94.3%	.481

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 Krippendorff's Alpha measure scorer agreement. The UNCW benchmarks is .67 for Krippendorff's alpha. See Appendix B of the General Education Assessment 2013 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, there are no dimensions of the rubric that meet these standards. Looking at percent agreement plus adjacent (that is, the scores that were within one level of each other), we find that all dimensions had greater than 80% of scores within one level of each other.

Scorer Feedback on Rubric and Required Assumptions

Scorer opinion about the fit of rubric dimensions to the assignments was mixed. Only for CT1 Explanation of the Issues and CT2a Evidence – Selecting and Using did all scorers agree that the dimension fit the assignment, either with or without guidelines. For CT3b Influence of Context and Assumptions – Context, CT4 Student's Position, and CT5 Conclusions and Related Outcomes, the majority of scorers respond that these dimensions fit with guidelines. CT2b Evidence – Examining Viewpoint of Author and CT3a Influence of Context and Assumptions – Assumptions had most scorers responding that these dimensions did not fit the assignment, even with guidelines. Overall for CT2b through CT5, responses were mixed and varied from "fit well" to "did not fit, even with guidelines present".

In response to the open-ended questions about the rubric dimensions that did not fit the assignment, three major themes were evident. For the first theme, many scorers agreed that CT2b and CT3a did not fit the assignment and were unclear as to whose viewpoints were being addressed and how to analyze/interpret the viewpoints of student. The second theme addressed CT4 and CT5 and how content for these dimensions was not explicitly asked for or required in a large portion of the assignments. The last theme addressed the fact that the assignments given did not give students room to explore or encourage critical thinking. Some of the assignments dealt with data or concrete evidence so there was little to no room for critical inquiry or for examining the author's viewpoint.

Scorers also provided feedback on specific quality criteria that were problematic and on what improvements could be made to the rubric. Formatting and quality of the assignment or prompt was the biggest problem noted by scorers. Many scorers felt that the assignments did not ask for analysis, had constrictive word limits, and did not ask for performance beyond the benchmark level (1). Other feedback indicates that scorers felt that prompts did not clarify expectations and explain how to answer questions using critical thinking. One scorer brought up the quality of one assignment and how it may have been dampened due to the fact that it was an extra credit assignment. For improvements, one scorer recommended making the formatting of the rubric clearer and better organized by using bullets to highlight aspects required consistently. Other scorers suggested making clearer CT2b Examining the Viewpoint of the Author, CT4 Student's Position, and CT5 Conclusions and Outcomes.

The last section of open-ended questions involved the specific language used in the instructions and how it affected student responses. The first question was on overall extent and specificity of instructions and how they affected student responses. Responses were mixed fairly evenly with some scorers expressing that that instructions were fairly clear and well written and specific. The other half felt that instructions needed more examples of analysis, that the word limit prohibited expansion on student ideas, and that specificity and instructions did not go far enough and needed to be more explicit. The last question concerned the parts of the assignment that were the most effective in eliciting evidence of student critical thinking. Scorers felt words phrases such as, "explain fully," "support your analysis," "provide reasons to support your ideas," and "why" statements helped elicit interpretation and encouraged students to support their answers.

Discussion

Table 4.2 shows the percent of work products scored at a level two or higher and the percent of work products scored at a level three or higher for each dimension.

Table 4.2 Critical Thinking Percent of Sample Scored at or Above Benchmark

Dimension	% of Work Products Scored Two or Higher	% of Work Products Scored Three or Higher
CT 1 Explanation of Issues	57.2%	14.3%
CT2a Evidence: Selecting and Using	53.8%	11.5%
CT2b Evidence: Critically Examining for Viewpoint	34.8%	10.6%
CT3a Influence of Assumptions	32.0%	5.1%
CT3b Influence of Context	35.4%	8.0%
CT4 Student's Position	39.0%	5.0%
CT5 Conclusions and Related Outcomes	35.4%	6.2%

All work products scored on the Critical Thinking rubric were collected from lower-division courses. The benchmark achievement level for any lower-division course is a level two. For two dimensions, CT1, CT2a Evidence: Selecting and Using, more than half of the work products scored at a level two or better. For all other dimensions, fewer than two in five work products scored at a level two or above. The percentage of work products scored a level three or better ranged between 5% and 14%. The performance expectations for first- and second-year students is a 2. Although 37.6% of students had attained greater than 60 credits before taking this course, the nature of the assignments, all from 100- and 200-level courses, likely played a part in the performance of those students as well as the freshman and sophomores (recall that there was no correlation between scores and credit hours completed).

Of particular importance to note are the two dimensions with the highest percentages of work products scoring a two or better: CT1 Explanation of Issues and CT2a: Evidence: Selecting and Using. In addition to being the highest-scoring dimensions, these two dimensions are also the only two that scorers agreed fit the assignments. A recurrent theme in scorers' feedback about the assignments was that the instructions limited the students' achievement on the rubric dimensions. Scorers pointed out that students were not always given instructions to critically explore ideas or to critically examine the viewpoints of other authors. This feedback, coupled with the results for dimensions CT1 and CT2a, suggest that the assignments might be limiting achievement on the critical thinking dimensions.

These findings, that students seem to be stronger on CT1 and CT2a, are aligned with those from an earlier study, the Fall 2010 Critical Thinking study. Also mirroring results from that study is the finding that out-of-class assignments tended to score higher than in-class assignments such as essay questions. These results, combined with the scorer feedback, indicate that students need more practice in how to critically examine the viewpoints in their resources, identify and analyze context and assumptions, clearly state their position in their writing, and tie their conclusions to the information presented in their writing.

APPENDIX 4.A CRITICAL THINKING RUBRIC

CRITICAL THINKING VALUE RUBRIC (AAC&U)

Modification 2: May 2013 UNCW

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 in10ded 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

Critical thinking is a habit of mind characterized by the comprehensive exploration of issues, ideas, artifacts, and events before accepting or formulating an opinion or conclusion.

Framing Language

This rubric is designed to be transdisciplinary, reflecting the recognition that success in all disciplines requires habits of inquiry and analysis that share common attributes. Further, research suggests that successful critical thinkers from all disciplines increasingly need to be able to apply those habits in various and changing situations encountered in all walks of life.

This rubric is designed for use with many different types of assignments and the suggestions here are not an exhaustive list of possibilities. Critical thinking can be demonstrated in assignments that require students to complete analyses of text, data, or issues. Assignments that cut across presentation mode might be especially useful in some fields. If insight into the process components of critical thinking (e.g., how information sources were evaluated regardless of whether they were included in the product) is important, assignments focused on student reflection might be especially illuminating.

Glossary

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

- Assumptions: Ideas, conditions, or beliefs (of10 implicit or unstated) that are "taken for granted or accepted as true without proof." (quoted from www.dictionary.reference.com/browse/assumptions)
- Context: The historical, ethical. political, cultural, environmental, or circumstantial settings or conditions that influence and complicate the consideration of any issues, ideas, artifacts, and events.

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

	Benchmark	Milestones		Capstone	
	1	2	3	4	Score
1. Explanation of Issues	Issue/problem to be considered critically is stated without clarification or description.	Issue/problem to be considered critically is stated but description leaves some aspects unexplored.	Issue/problem to be considered critically is stated, described, and clarified so that understanding is not seriously impeded by omissions.	Issue/problem to be considered critically is stated clearly and described comprehensively, delivering all relevant information necessary for full understanding.	
2. Evidence *a. Selecting and using information*	Information is taken from source(s) without any interpretation/evaluation.	Information is taken from source(s) with some interpretation/evaluation, but not enough to develop a coherent analysis or synthesis.	Information is taken from source(s) with enough interpretation/evaluation to develop a coherent analysis or synthesis.	Information is taken from source(s) with enough interpretation/evaluation to develop a comprehensive analysis or synthesis.	
b. Critically examining evidence for viewpoint	Viewpoints of authors are taken as fact, without question.	Viewpoints of authors are taken as mostly fact, with little questioning.	Viewpoints of authors are subject to questioning.	Viewpoints of authors are questioned thoroughly.	
3. Influence of context and assumptions *a. Assumptions*	Shows an emerging awareness of present assumptions (own or others').	Questions some assumptions. May be more aware of others' assumptions than one's own (or vice versa).	Identifies own and others' assumptions when presenting a position.	Thoroughly (systematically and methodically) analyzes own and others' assumptions when presenting a position.	
b. Context	Begins to identify some contexts when presenting a position.	Identifies several relevant contexts when presenting a position.	Identifies several relevant contexts and discusses at least some aspects of their interconnectedness.	Carefully evaluates the relevance of context when presenting a position.	
4. Student's position *(position, perspective, thesis, or hypothesis)*	Specific position is stated, but is simplistic and obvious.	Specific position acknowledges different sides of an issue.	Specific position takes into account the complexities of an issue. Others' points of view are acknowledged within position.	Specific position is imaginative, taking into account the complexities of an issue. Limits of position are acknowledged. Others' points of view are synthesized within position.	
5. Conclusions and related outcomes *(implications and consequences)*	Conclusion is inconsistently tied to some of the information discussed; related outcomes are oversimplified.	Conclusion is logically tied to information (because information is chosen to fit the desired conclusion); some related outcomes are identified.	Conclusion is logically tied to a range of information, including opposing viewpoints; related outcomes are identified clearly.	Conclusions and related outcomes are logical and reflect student's informed evaluation and ability to place evidence and perspectives discussed in priority order.	

APPENDIX 4.B CORRELATIONS BETWEEN CRITICAL THINKING DIMENSIONS

Spearman rho Rank Order Correlation Coefficients

Correlations

			CT1	CT2a	CT2b	CT3a	CT3b	CT4	CT5
Spearman's rho	CT1	Correlation Coefficient		.654**	.564**	.602**	.622**	.646**	.585**
		N		175	161	175	175	159	161
	CT2a	Correlation Coefficient	.654**		.575**	.667**	.643**	.520**	.547**
		N	175		161	175	175	159	161
	CT2b	Correlation Coefficient	.564**	.575**		.735**	.717**	.541**	.563**
		N	161	161		161	161	145	161
	CT3a	Correlation Coefficient	.602**	.667**	.735**		.743**	.623**	.582**
		N	175	175	161		175	159	161
	CT3b	Correlation Coefficient	.622**	.643**	.717**	.743**		.713**	.625**
		N	175	175	161	175		159	161
	CT4	Correlation Coefficient	.646**	.520**	.541**	.623**	.713**		.544**
		N	159	159	145	159	159		145
	CT5	Correlation Coefficient	.585**	.547**	.563**	.582**	.625**	.544**	
		N	161	161	161	161	161	145	

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

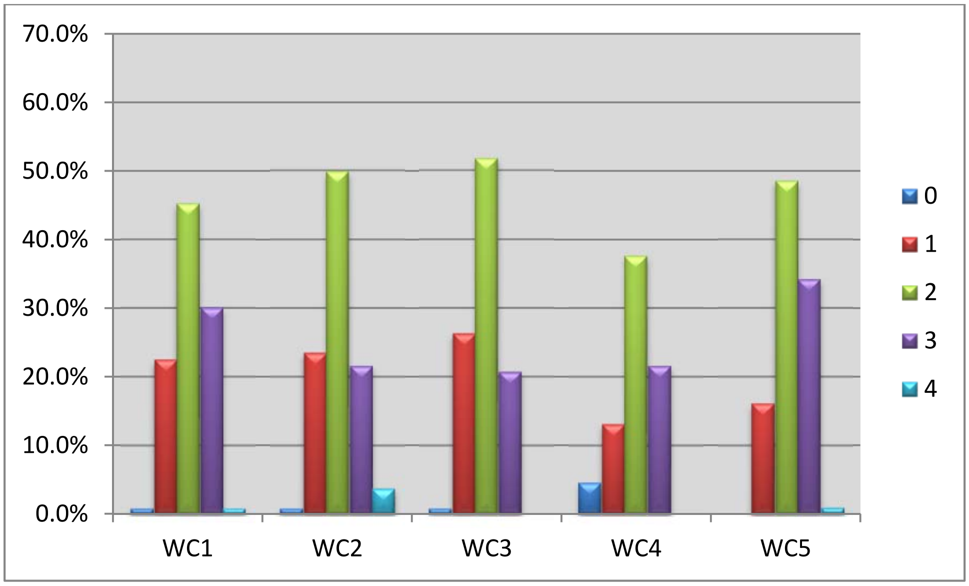
5. THOUGHTFUL EXPRESSION (WRITTEN)

The UNCW Thoughtful Expression (Written) learning goal is for students to demonstrate an ability to express meaningful ideas in writing. For purposes of this Learning Goal, "thoughtful expression is the ability to communicate meaningful ideas in an organized, reasoned and convincing manner. Thoughtful expression involves a purpose responsive to an identified audience, effective organization, insightful reasoning and supporting detail, style appropriate to the relevant discipline, purposeful use of sources and evidence, and error-free syntax and mechanics" (UNCW Learning Goals, 2011). The VALUE Written Communication rubric contains five dimensions that are aligned with the UNCW description of Thoughtful Expression. The Thoughtful Expression (Written) rubric can be found in Appendix 5.A at the end of this chapter. In this study, work was sampled from the Writing Intensive and the Aesthetic, Interpretive, and Literary Perspective components of University Studies.

SUMMARY OF SCORES BY DIMENSION

Eight faculty scorers scored 187 work products from eight assignments across six courses from the Spring 2013 semester: ACG445, FST110, MUS115, NSG415, SED372, and THR121. ACG 445, NSG 415, and SED 372 are in the Writing Intensive component of University Studies. FST 110, MUS 115, and THR 121 are Aesthetic, Interpretive, and Literary Perspective courses. The types of assignments consisted of research papers, critique papers, and projects, some completed as groups and some individually. Forty-two work products (22.5%) were scored by multiple scorers. Figures 5.1.a and 5.1.b provide the score distributions for each dimension for work products that were scored on that dimension (i.e., work products with blank scores are not included).

THOUGHTFUL EXPRESSION (WRITTEN) RESULTS BY DIMENSION FOR APPLICABLE SCORES ONLY, LOWER-DIVISION COURSES



0	0.9%	0.9%	0.9%	4.7%	0.0%
1	22.6%	23.6%	26.4%	13.2%	16.2%
2	45.3%	50.0%	51.9%	37.7%	48.6%
3	30.2%	21.7%	20.8%	21.7%	34.3%
4	0.9%	3.8%	0.0%	0.0%	1.0%
25th %tile	2	1.75	1	2	2
50th %tile	2	2	2	2	2
75th %tile	3	3	2	3	3
Mode	2	2	2	2	2
N	106	106	106	82	105

Figure 5.1.a Distribution of Scores for Thoughtful Expression (Written) in Lower-Division Courses, Applicable Scores Only

RESULTS BY DIMENSION

WC1 Context of and Purpose for Writing

Fewer than one in 10 work products demonstrated complete lack of attention to context, audience, purpose and to the assigned task (score of 0). Slightly over one in five work products demonstrated minimal attention to context, audience, purpose, and to the assigned task (scores of 1). Just under half of the work products demonstrated awareness of the context, audience, purpose, and assigned task (scores of 2). Three out of 10 work products demonstrated adequate consideration of context, audience, and purpose, and a clear focus on the assigned task (scores of 3). Nine out of 10 work products demonstrated a thorough understanding of context, audience, and purpose that was responsive to the assigned task and focused all elements of the work (scores of 4).

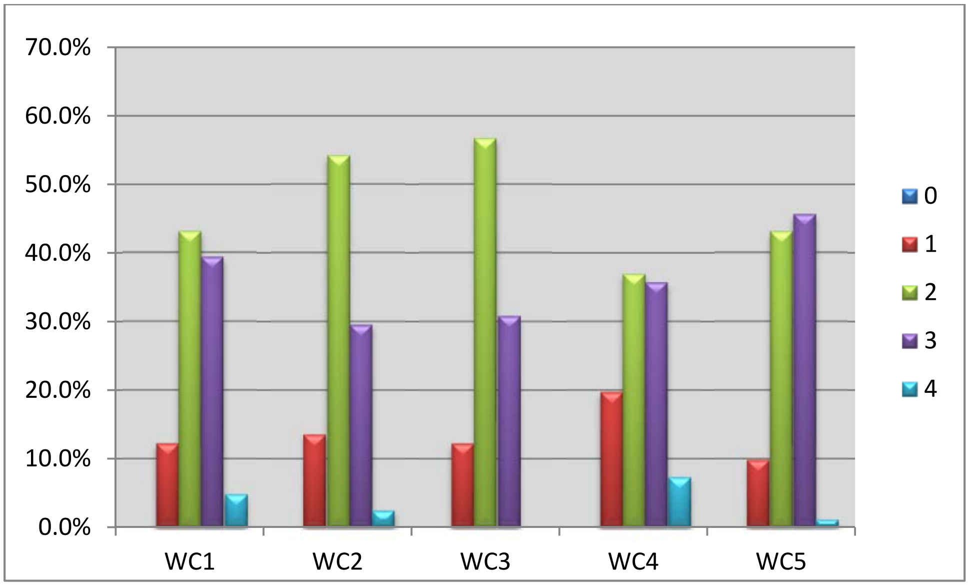
WC2 Content Development
Fewer than one in 10 work products demonstrated no content development (score of 0). Slightly under one-fourth of work products used appropriate and relevant content to develop simple ideas in some parts of the work (scores of 1). Half of the work products used appropriate and relevant content to develop and explore ideas through the work (scores of 2). Just over two out of 10 work products used appropriate, relevant and compelling content to explore ideas within the context of the discipline (scores of 3). One in 30 work products used appropriate, relevant, and compelling content to illustrate mastery of the subject, conveying the writer's understanding, and shaping the whole work (scores of 4).

WC3 Genre and Disciplinary Conventions
Less than one in 10 work products failed to show an attempt to use a consistent system for organization and presentation (score of 0). Slightly more than one in four work products demonstrated an attempt to use a consistent system for basic organization and presentation (scores of 1). Over half of the work products followed expectations appropriate to the specific writing task for basic organization, content, and presentation (scores of 2). Two out of 10 work products demonstrated consistent use of important conventions particular to the writing task, including stylistic choices (scores of 3). No work products received a score of four on this dimension.

WC4 Sources and Evidence
One in twenty-five work products demonstrated no attempt to use sources to support ideas (score of 0). Just over one in eight work products demonstrated an attempt to use sources to support ideas (scores of 1). Over one-third of work products demonstrated an attempt to use credible and/or relevant sources to support ideas that were appropriate to the task (scores of 2). Just over two in 10 work products demonstrated consistent use of credible, relevant sources to support ideas (scores of 3). No work products demonstrated skill use of high-quality, credible, relevant sources to develop ideas that are appropriate for the discipline (scores of 4).

WC5 Control of Syntax and Mechanics
No work products failed to meet the level 1 benchmark (score of 0). Approximately one in six work products used language that sometimes impeded meaning because of errors in usage (score of 1). Just under half of the work products used language that generally conveyed meaning with clarity, although writing included some errors (score of 2). Just over one in three work products used straightforward language that generally conveyed meaning, with few errors (score of 3). One in 100 work products used graceful language that skillfully communicated meaning with clarity and fluency, with virtually no errors (scores of 4).

Thoughtful Expression (Written) Results by Dimension for Applicable Scores Only, Upper-division Courses



0	0.0%	0.0%	0.0%	0.0%	0.0%
1	12.3%	13.6%	12.3%	19.8%	9.9%
2	43.2%	54.3%	56.8%	37.0%	43.2%
3	39.5%	29.6%	30.9%	35.8%	45.7%
4	4.9%	2.5%	0.0%	7.4%	1.2%
25th %tile	2	2	2	2	2
50th %tile	2	2	2	2	2
75th %tile	3	3	3	3	3
Mode	2	2	2	2	3
N	81	81	81	81	81

Figure 5.1.b Distribution of Scores for Thoughtful Expression (Written) in Upper-Division Courses, Applicable Scores Only

Results by Dimension

WC1 Context of and Purpose for Writing

No student work products demonstrated complete lack of attention to context, audience, purpose and to the assigned task (score of 0). One in eight work products demonstrated minimal attention to context, audience, purpose, and to the assigned task (scores of 1). Just less than half of the work products demonstrated awareness of the context, audience, purpose, and assigned task (scores of 2). Slightly less than four out of 10 work products demonstrated adequate consideration of context, audience, and purpose, and a clear focus on the assigned task (scores of 3). Five in 100 work products demonstrated a thorough understanding of context, audience, and purpose that was responsive to the assigned task and focused all elements of the work (scores of 4).

WC2 Content Development
None of the student work products demonstrated lack of content development (score of 0). Just less than one in seven work products used appropriate and relevant content to develop simple ideas in some parts of the work (scores of 1). Just over half of the work products used appropriate and relevant content to develop and explore ideas through the work (scores of 2). One out of 10 work products used appropriate, relevant and compelling content to explore ideas within the context of the discipline (scores of 3). Finally, one out of 40 work products used appropriate, relevant, and compelling content to illustrate mastery of the subject, conveying the writer's understanding, and shaping the whole work (scores of 4).

WC3 Genre and Disciplinary Conventions
No work products failed to show an attempt to use a consistent system for organization and presentation (score of 0). One out of eight work products demonstrated an attempt to use a consistent system for basic organization and presentation (scores of 1). Just over half of the work products followed expectations appropriate to the specific writing task for basic organization, content, and presentation (scores of 2). One out of 10 work products demonstrated consistent use of important conventions particular to the writing task, including stylistic choices (scores of 3). No work products received a score of four on this dimension.

WC4 Sources and Evidence
No student work products failed to demonstrate an attempt to use sources to support ideas. Two out of 10 work products demonstrated an attempt to use sources to support ideas (scores of 1). Just less than four in 10 work products demonstrated an attempt to use credible and/or relevant sources to support ideas that were appropriate to the task (scores of 2). Just over one third of the work products demonstrated consistent use of credible, relevant sources to support ideas (scores of 3). One in 14 work products demonstrated skill use of high-quality, credible, relevant sources to develop ideas that are appropriate for the discipline (scores of 4).

WC5 Control of Syntax and Mechanics
No work products failed to meet the level 1 benchmark (score of 0). Approximately one out of 10 work products used language that sometimes impeded meaning because of errors in usage (score of 1). Over four in 10 work products used language that generally conveyed meaning with clarity, although writing included some errors (score of 2). Just under half of the work products used straightforward language that generally conveyed meaning, with few errors (score of 3). One in 100 work products used graceful language that skillfully communicated meaning with clarity and fluency, with virtually no errors (scores of 4).

CORRELATION BETWEEN DIMENSIONS
All dimensions were correlated at the .001 level of significance. See Appendix 5.B for the complete presentation of dimension correlation coefficients.

Demographic and Preparedness Findings

There were some statistically significant difference between the means, medians, and the score distributions of males and females, transfer vs. UNCW-start students, and honors vs. non-honors students. Females scored higher than did males on WC4. UNCW-start students scored higher on WC5 than transfer students. Honors students scored lower on two dimensions, WC4 and WC5, than non-honors students. It should be noted that the number of Honors students in the sample was very small at 15.

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 (22.0% of the sample), 31 – 60 credit hours (26.8% of the sample), 61 – 90 (14.3% of the sample), and over 90 credit hours (36.9% 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 for any of the Thoughtful Expression (Written) dimensions. Looking at Spearman rho correlation coefficients, the number of total hours completed was not significantly correlated with any of the Thoughtful Expression (Written) dimensions.

There were some significant correlations with GPA and SAT-Verbal scores and the Thoughtful Expression (Written) dimensions. GPA was positively correlated with three dimensions: WC1 (.157*), WC2 (.197*), and WC3 (.156*). WC1, WC3 and WC5 were positively correlated with SAT-Verbal (.229*, .198* and .412**, respectively). SAT-Math scores were positively correlated with WC5 (.180*). There were no significant correlations between ACT scores and the Thoughtful Expression (Written) dimension scores.

Comparisons between upper- and lower-division courses and assignment types

The assignments, instructional setting, and instructor type of the work products collected were varied. Statistical analyses indicated differences between the scores of some of these groups. Assignments completed in classes taught by tenure-track faculty scored statistically higher on WC1, WC3, and WC5. There was no statistical difference between the scores on group assignments vs. independent assignments. There was no difference in the scores of work products completed for online courses vs. traditional courses. There were statistical differences seen in the work products from the different course levels on WC1, WC3, and WC4, with the 300-level coursework scoring higher on all three dimensions, followed by the 400-level work and then the 100-level work.

Comparisons between University Studies components

To assess thoughtful expression (written), work was sampled and scored from two components of University Studies: Aesthetic, Interpretive, and Literary Perspectives (n=106) and Writing Intensive (n=81). There were statistical differences between the scores of work sampled from these two components for dimensions WC1, WC3, and WC4. For these three dimensions, work sampled from Writing Intensive courses was scored higher. It should be noted that all the AILP courses were 100-level courses, whereas the WI courses were from the 300- and 400-levels.

Interrater Reliability

For each group of scorers, there were a number of common papers scored so that interrater reliability could be assessed (42 or 22.5%). Table 5.1 shows the reliability measures for Information Literacy.

Table 5.1 Interrater Reliability for Thoughtful Expression (Written)

Dimension	N	Percent Agreement	Plus Percent Adjacent	Krippendorff's Alpha
WC1 Context of and Purpose for Writing	42	57.1%	92.9%	.530
WC2 Content Development	42	61.9%	100.0%	.746
WC3 Genre and Disciplinary Conventions	42	50.0%	95.2%	.559
WC4 Sources and Evidence	37	56.8%	97.3%	.761
WC5 Control of Syntax and Mechanics	42	57.1%	95.2%	.519

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 Krippendorff's Alpha measure scorer agreement. The UNCW benchmarks are .67 for Krippendorff's alpha. See Appendix B of the General Education Assessment 2013 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, there are two dimensions of the rubric that meet these standards, WC2 and WC4. Looking at percent adjacent (that is, the scores that were within one level of each other), we find that all dimensions had greater than 92% of scores within one level of each other.

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 guidelines”, with more scorers agreeing that the rubric “fit well” on each dimension except for WC3 Genre and Disciplinary Conventions. For two dimensions, WC3 Genre and Disciplinary Conventions and WC4 Sources and Evidence, scorer opinion was mixed between “fit well”, “fit with guidelines”, and “did not fit”.

In response to the open-ended questions about the rubric dimensions that did not fit the assignment, the majority of scorers addressed problems with the assignment and how WC3 Genre and Disciplinary Conventions and WC4 Sources of Evidence did not fit. Comments for the dimension WC3 included that the needed information was not presented in the assignment and must have been given prior which would leave the scorer to make assumptions, and that more specific instructions were needed. Comments for the dimension WC4 included that most assignments did not require references or citations which made it difficult to assess this dimension. Two scorers chose to focus on incorporation of details from the student’s observation instead of scoring on sources because they were not required in the assignment. Other scorers commented on the difficulty to assess WC5 Control of Syntax and Mechanics with literary reviews and difficulty with WC2 Content Development when it is not addressed in the assignment and only seen in demonstration by the student.

Scorers also provided feedback on which specific quality criteria were problematic and what improvements could be made to the rubric. The majority of problems were seen in the differentiation between milestone levels in all of the dimensions, with WC5 Control of Syntax and Mechanics being named by multiple scorers as particularly problematic in this sense. Scorers felt like more explicit definition of the dimension levels needed to be emphasized with more examples of the difference, for instance, between terms such as “some errors” and “few errors” or “lacks clarity” and “impedes meaning”.

The last section of open-ended questions involved the specific language used in the instruction and how it affected student responses. The first question was on overall extent and specificity of instructions. The majority of scorers commented that the instructions were well organized, very specific and clear. On the other hand they noted that the instructions also tended to be very brief and generic which lead to basic summarizations. Some of the assignments did not ask for formal sources of evidence with one scorer commenting that, “too many opinions were given as basis of analysis”. Student responses tended to be concise and similar with some of the scorers commenting that they were “generically organized essays” and “everyone used some of the same phrases and sources to respond/format their papers”. The last question concerned the parts of the assignment that were particularly most effective in eliciting evidence of student writing communication. Scorers felt that statements such as, “critically analyze”, “include rationale”, and “terms and concepts” helped students form their responses. Verbs such as, “summarize”, “evaluate”, “critique”, and “elaborate” gave students direction in their writings.

Discussion

Table 5.2 shows the percent of work products scored at a level two or higher and the percent of work products scored at a level three or higher for each dimension. Level two is the benchmark for lower-division general education courses, and level three is the benchmark for upper-division courses.

Table 5.2 Thoughtful Expression (Written) Percent of Sample Scored at or Above 2 and 3

Lower-Division Courses		
Dimension	% of Work Products Scored Two or Higher	% of Work Products Scored Three or Higher
WC1 Context of and Purpose for Writing	76.4%	31.3%
WC2 Content Development	75.5%	25.5%
WC3 Genre and Disciplinary Conventions	72.7%	20.8%
WC4 Sources and Evidence	59.4%	21.7%
WC5 Control of Syntax and Mechanics	83.9%	35.3%
Upper-Division Courses		
Dimension	% of Work Products Scored Two or Higher	% of Work Products Scored Three or Higher
WC1 Context of and Purpose for Writing	87.6%	44.4%
WC2 Content Development	86.4%	32.1%
WC3 Genre and Disciplinary Conventions	87.7%	30.9%
WC4 Sources and Evidence	80.2%	43.2%
WC5 Control of Syntax and Mechanics	90.1%	46.9%

For work products from lower-division courses, the benchmark achievement level is set at a score of two. For all dimensions, a majority of student work collected from 100- and 200- level courses scored a two or higher. The benchmark level graduating seniors is a 4. Scores 3 and above are reported because the quality criteria describe solid performance and because not all students assessed were graduating at the end of the semester in which the work was sampled. For no dimension did a majority percentage of work collected from upper-division courses meet the level three.

For both upper- and lower-division courses, WC1 and WC5 had more papers meeting or exceeding the benchmark than then other dimensions. It is not surprising that WC1 and WC5 were the higher-scoring dimensions. Dimensions that can be construed as drawing upon more sophisticated skills, such as the use of content to develop one's own ideas, using sources to support one's on position, and using disciplinary conventions, scored lower. WC4 was not scored for all the work products; it was deemed "not applicable" for 24 (12.8%) of the work products scored.

The classes from which the student work was selected included 100-level (56.6%), 300-level (16.6%), and 400-level (26.8%) courses. Though there was no meaningful difference in the scores related to the number of credit hours completed by students, there were statistical differences seen in the work products from the different course levels on three dimensions, with the 300-level work scoring higher than the 400- or 100-level work. This suggests that the student achievement on the Written Communication rubric may be related less to experience than to the particular assignment.

APPENDIX 5.A THOUGHTFUL EXPRESSION (WRITTEN) RUBRIC

WRITTEN COMMUNICATION VALUE RUBRIC

for more information, please contact value@aacu.org

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

Written communication is the development and expression of ideas in writing. Written communication involves learning to work in many genres and styles. It can involve working with many different writing technologies, and mixing texts, data, and images. Written communication abilities develop through iterative experiences across the curriculum.

Framing Language

This writing rubric is designed for use in a wide variety of educational institutions. The most clear finding to emerge from decades of research on writing assessment is that the best writing assessments are locally determined and sensitive to local context and mission. Users of this rubric should, in the end, consider making adaptations and additions that clearly link the language of the rubric to individual campus contexts.

This rubric focuses assessment on how specific written work samples or collections of work respond to specific contexts. The central question guiding the rubric is "How well does writing respond to the needs of audience(s) for the work?" In focusing on this question the rubric does not attend to other aspects of writing that are equally important: issues of writing process, writing strategies, writers' fluency with different modes of textual production or publication, or writer's growing engagement with writing and disciplinarity through the process of writing.

Evaluators using this rubric must have information about the assignments or purposes for writing guiding writers' work. Also recommended is including reflective work samples of collections of work that address such questions as: What decisions did the writer make about audience, purpose, and genre as s/he compiled the work in the portfolio? How are those choices evident in the writing -- in the content, organization and structure, reasoning, evidence, mechanical and surface conventions, and citational systems used in the writing? This will enable evaluators to have a clear sense of how writers understand the assignments and take it into consideration as they evaluate

The first section of this rubric addresses the context and purpose for writing. A work sample or collections of work can convey the context and purpose for the writing tasks it showcases by including the writing assignments associated with work samples. But writers may also convey the context and purpose for their writing within the texts. It is important for faculty and institutions to include directions for students about how they should represent their writing contexts and purposes.

Faculty interested in the research on writing assessment that has guided our work here can consult the National Council of Teachers of English/Council of Writing Program Administrators' White Paper on Writing Assessment (2008; www.wpacouncil.org/whitepaper) and the Conference on College Composition and Communication's Writing Assessment: A Position Statement (2008; www.ncte.org/cccc/resources/positions/123784.htm)

Glossary

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

- Content Development: The ways in which the text explores and represents its topic in relation to its audience and purpose.
- Context of and purpose for writing: The context of writing is the situation surrounding a text: who is reading it? who is writing it? Under what circumstances will the text be shared or circulated? What social or political factors might affect how the text is composed or interpreted? The purpose for writing is the writer's intended effect on an audience. Writers might want to persuade or inform; they might want to report or summarize information; they might want to work through complexity or confusion; they might want to argue with other writers, or connect with other writers; they might want to convey urgency or amuse; they might write for themselves or for an assignment or to remember.
- Disciplinary conventions: Formal and informal rules that constitute what is seen generally as appropriate within different academic fields, e.g. introductory strategies, use of passive voice or first person point of view, expectations for thesis or hypothesis, expectations for kinds of evidence and support that are appropriate to the task at hand, use of primary and secondary sources to provide evidence and support arguments and to document critical perspectives on the topic. Writers will incorporate sources according to disciplinary and genre conventions, according to the writer's purpose for the text. Through increasingly sophisticated use of sources, writers develop an ability to differentiate between their own ideas and the ideas of others, credit and build upon work already accomplished in the field or issue they are addressing, and provide meaningful examples to readers.
- Evidence: Source material that is used to extend, in purposeful ways, writers' ideas in a text.
- Genre conventions: Formal and informal rules for particular kinds of texts and/or media that guide formatting, organization, and stylistic choices, e.g. lab reports, academic papers, poetry, webpages, or personal essays.
- Sources: Texts (written, oral, behavioral, visual, or other) that writers draw on as they work for a variety of purposes -- to extend, argue with, develop, define, or shape their ideas, for example.

WRITTEN COMMUNICATION VALUE RUBRIC

for more information, please contact value@aacu.org

Definition

Written communication is the development and expression of ideas in writing. Written communication involves learning to work in many genres and styles. It can involve working with many different writing technologies, and mixing texts, data, and images. Written communication abilities develop through iterative experiences across the curriculum.

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

	Benchmark	Milestones		Capstone	Score
	1	2	3	4	
1. Context of and Purpose for Writing *Includes considerations of audience, purpose, and the circumstances surrounding the writing task(s).*	Demonstrates minimal attention to context, audience, purpose, and to the assigned tasks(s) (e.g., expectation of instructor or self as audience).	Demonstrates awareness of context, audience, purpose, and to the assigned tasks(s) (e.g., begins to show awareness of audience's perceptions and assumptions).	Demonstrates adequate consideration of context, audience, and purpose and a clear focus on the assigned task(s) (e.g., the task aligns with audience, purpose, and context).	Demonstrates a thorough understanding of context, audience, and purpose that is responsive to the assigned task(s) and focuses all elements of the work.	
2. Content Development	Uses appropriate and relevant content to develop simple ideas in some parts of the work.	Uses appropriate and relevant content to develop and explore ideas through most of the work.	Uses appropriate, relevant, and compelling content to explore ideas within the context of the discipline and shape the whole work.	Uses appropriate, relevant, and compelling content to illustrate mastery of the subject, conveying the writer's understanding, and shaping the whole work.	
3. Genre and Disciplinary Conventions *Formal and informal rules inherent in the expectations for writing in particular forms and/or academic fields.*	Attempts to use a consistent system for basic organization and presentation.	Follows expectations appropriate to a specific discipline and/or writing task(s) for basic organization, content, and presentation	Demonstrates consistent use of important conventions particular to a specific discipline and/or writing task(s), including organization, content, presentation, and stylistic choices	Demonstrates detailed attention to and successful execution of a wide range of conventions particular to a specific discipline and/or writing task(s) including organization, content, presentation, formatting, and stylistic choices	
4. Sources and Evidence	Demonstrates an attempt to use sources to support ideas in the writing.	Demonstrates an attempt to use credible and/or relevant sources to support ideas that are appropriate for the discipline and genre of the writing.	Demonstrates consistent use of credible, relevant sources to support ideas that are situated within the discipline and genre of the writing.	Demonstrates skillful use of high-quality, credible, relevant sources to develop ideas that are appropriate for the discipline and genre of the writing	
5. Control of Syntax and Mechanics	Uses language that sometimes impedes meaning because of errors in usage.	Uses language that generally conveys meaning to readers with clarity, although writing may include some errors.	Uses straightforward language that generally conveys meaning to readers. The language in the portfolio has few errors.	Uses graceful language that skillfully communicates meaning to readers with clarity and fluency, and is virtually error-free.	

Appendix 5.B Correlations between Thoughtful Expression (Written) Dimensions

Spearman rho Rank Order Correlation Coefficients

Correlations

			WC1	WC2	WC3	WC4	WC5
Spearman's rho	WC1	Correlation Coefficient		.592**	.639**	.530**	.573**
		N		187	187	163	186
	WC2	Correlation Coefficient	.592**		.646**	.508**	.599**
		N	187		187	163	186
	WC3	Correlation Coefficient	.639**	.646**		.434**	.556**
		N	187	187		163	186
	WC4	Correlation Coefficient	.530**	.508**	.434**		.486**
		N	163	163	163		163
	WC5	Correlation Coefficient	.573**	.599**	.556**	.486**	
		N	186	186	186	163	

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

6. 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). Speaking and listening proficiency were assessed in French and Spanish in Fall 2012. The rubrics used to assess these proficiencies were developed at UNCW are included in Appendix 6.A at the end of this chapter.

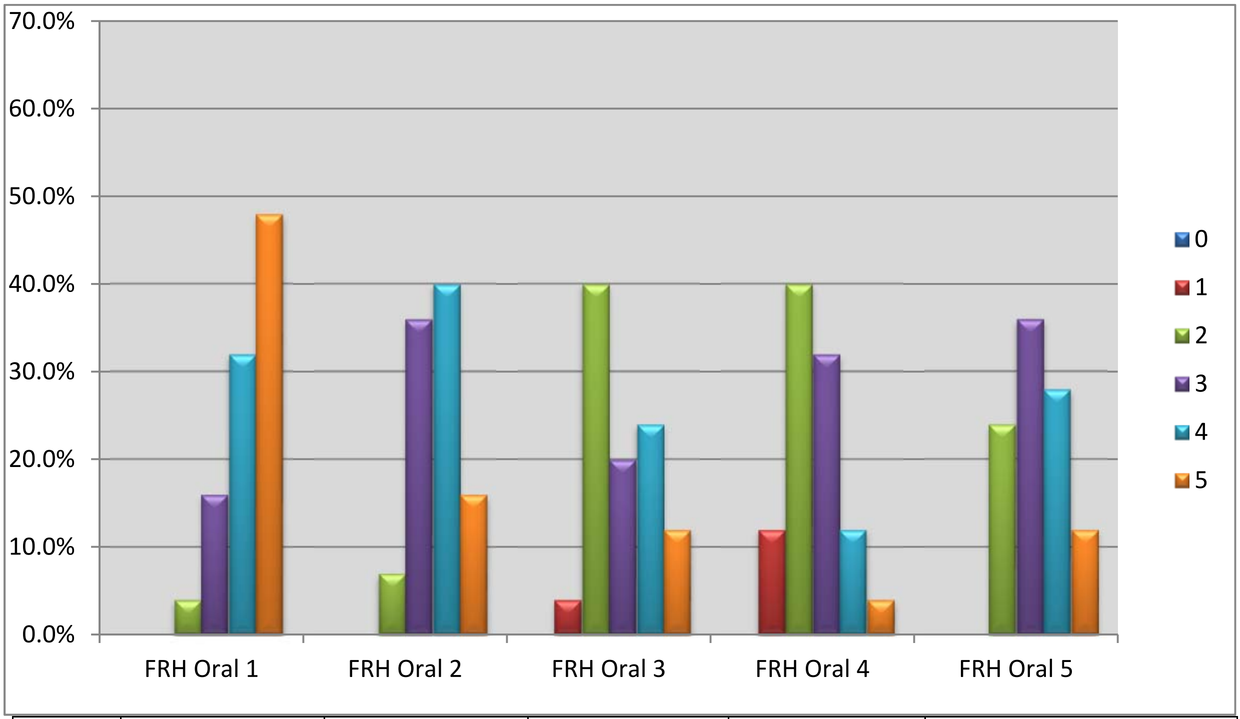
FRENCH SPEAKING AND LISTENING

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

SUMMARY OF SCORES BY DIMENSION

Two faculty scorers scored 25 student oral interview exams from one course from the Fall 2012 semester, FRH 201. 10 interviews (40.0%) were scored by both scorers. Figure 6.1 provides the score distributions for each dimension for the oral interviews that were scored on that dimension (interviews scored as not applicable [NA] are not included).

French Speaking and Listening Results by Dimension for Applicable Scores Only



0	0.0%	0.0%	0.0%	0.0%	0.0%
1	0.0%	0.0%	4.0%	12.0%	0.0%
2	4.0%	7.0%	40.0%	40.0%	24.0%
3	16.0%	36.0%	20.0%	32.0%	36.0%
4	32.0%	40.0%	24.0%	12.0%	28.0%
5	48.0%	16.0%	12.0%	4.0%	12.0%
25th %tile	4	3	2	2	2.5
50th %tile	4	4	3	2	3
75th %tile	5	4	4	3	4
Mode	5	4	2	2	3
N	25	25	25	25	25

Figure 6.1 Distribution of Scores for French Speaking and Listening, Applicable Scores Only

Results by Dimension

FRH Oral 1 Listening Comprehension

Scores on listening comprehension were highest of the five dimensions. No students scored a zero or one on this dimension. One in 25 students scored a two, indicating that the responses provided during the interview were mostly not appropriate. Four in 25 students scored a three on this dimension, indicating that responses were often inappropriate. Just under one-third of the students gave responses that were mostly appropriate (scores of four) and almost one half of the interviews scored indicated that students almost always provided appropriate responses (scores of five).

FRH Oral 2 Pronunciation

The second-highest scores were on the pronunciation dimension. Just as for dimension one, no students scored a zero or one on this dimension. Two students demonstrated severe interference from their native language in the pronunciation of the French words used in the interview (scores of two). Substantial interference from the native language was observed in nine of 25 interviews (scores of three). Two in five students received scores of four, demonstrating occasional interference from the native language. Showing little or no interference from the native language, four out of 25 students scored a five on this dimension.

FRH Oral 3 Vocabulary

No students scored a zero for this dimension. One student in 25 showed a lack of course-appropriate vocabulary and demonstrated substantial errors during the oral interview (score of one). Two in five students received scores of two, insufficiently using course-appropriate vocabulary with frequent errors. One in five students exhibited occasional use of course-appropriate vocabulary with more than occasional errors and received scores of three. Six in 25 students' interviews were given scores of four, indicating that there was frequent use of course-appropriate vocabulary with few errors. Finally, three out of 25 students showed extensive use of course-appropriate vocabulary, with almost no errors, and received scores of five.

FRH Oral 4 Grammar

No students scored a zero on the Grammar dimension. Three in 25 students scored a level one, meaning that the correct usage of grammar was almost non-existent during the interview. Two in five students had substantial grammatical errors that obscured the meaning of their statements (scores of two). One in five students scored a level three, indicating that there were several errors and some avoidance of structures during the oral interview. One or two significant errors were observed in three out of 25 students' interviews (scores of four). Finally, one in 25 students had no significant errors (scores of five).

FRH Oral 5 Fluency

The same as for dimensions FRH Oral 1 through FRH Oral 4, this dimension had no students scoring at a level zero. No students were assigned scores of one for this dimension, either. Six out of 25 students exhibited many pauses with communication breakdown (scores of two). Nine in 25 students were given scores of three, indicating that there was frequent hesitation, though no significant breakdown of communication, during their interviews. Demonstrating only slight hesitation with the natural pauses, seven students received scores of four. Three out of 25 students exhibited only natural pauses (scores of five).

Correlation between Dimensions

All dimension scores were correlated with each other at the .01 level of significance. The range of correlation coefficients was .573 to .767, with FRH Oral 4 Grammar and FRH Oral 3 Vocabulary having the highest correlation. See Appendix 6.B at the end of this chapter for a complete presentation of correlation coefficients.

Demographic and Preparedness Findings

There were no statistically significant differences between the means, medians, and the score distributions of males vs. females. When comparing the scores of transfer vs. freshman-start students, there was a statistical difference between the scores of these groups on FRH Oral 4 Grammar, with freshman-start students scoring higher on this dimension. 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 (28.0% of the sample), 31 – 60 credit hours (48.0% of the sample), 61 – 90 (20.0% of the sample), and over 90 credit hours (4.0% 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 for any of the French speaking and listening rubric dimensions. Looking at Spearman rho correlation coefficients, the number of total hours completed was negatively significantly correlated with FRH Oral 2 (-.523**), FRH Oral 3 (-.469*), FRH Oral 4 (-.457*), and FRH Oral 5 (-.425*).

GPA was significantly correlated with FRH Oral 5 (.705**). SAT-Verbal was significantly correlated with FRH Oral 2 (.669*) and FRH Oral 4(.564*), and SAT-Math was significantly correlated with FRH Oral 4(.541*).

Interrater Reliability

In some instances, multiple scorers rated the student interviews so that interrater reliability could be assessed (10 or 40%). Table 6.1 shows the reliability measures for French Writing.

Table 6.1 Interrater Reliability for French Speaking and Listening

Dimension	Percent Agreement	Plus Percent Adjacent	Krippendorff's Alpha
FRH Oral 1 Listening Comprehension	80.0%	100.0%	.866
FRH Oral 2 Pronunciation	90.0%	100.0%	.947
FRH Oral 3 Vocabulary	30.0%	90.0%	.670
FRH Oral 4 Grammar	50.0%	100.0%	.744
FRH Oral 5 Fluency	80.0%	100.0%	.935

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 Krippendorff's Alpha measure scorer agreement. The UNCW benchmarks are .67 for Krippendorff's alpha. See Appendix B of the General Education Assessment 2013 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, all dimensions met these standards. Looking at percent agreement plus adjacent (that is, the scores that were within one level of each other), we find that all dimensions had greater than 90% of scores in agreement or within one level of each other, with four of the five having 100% agreeing or adjacent scores.

Discussion

This was the first independent study using the French Speaking and Listening Assessment Rubric. Table 6.2 shows the percent of work products scored at a level 3 or higher for each dimension. Level 3 is the benchmark for proficiency.

Table 6.2 French Speaking and Listening Percent of Sample Scored at or above 3 and 4

Dimension	% of Work Products Scored 3 or higher	% of Work Products Scored 4 or higher
FRH Oral 1 Listening Comprehension	96.0%	80.0%
FRH Oral 2 Pronunciation	93.0%	56.0%
FRH Oral 3 Vocabulary	56.0%	36.0%
FRH Oral 4 Grammar	48.0%	16.0%
FRH Oral 5 Fluency	76.0%	40.0%

The results indicate Listening Comprehension is a relative strength and Vocabulary and Grammar are areas for improvement. Interrater reliability statistics indicate that these scores are 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.

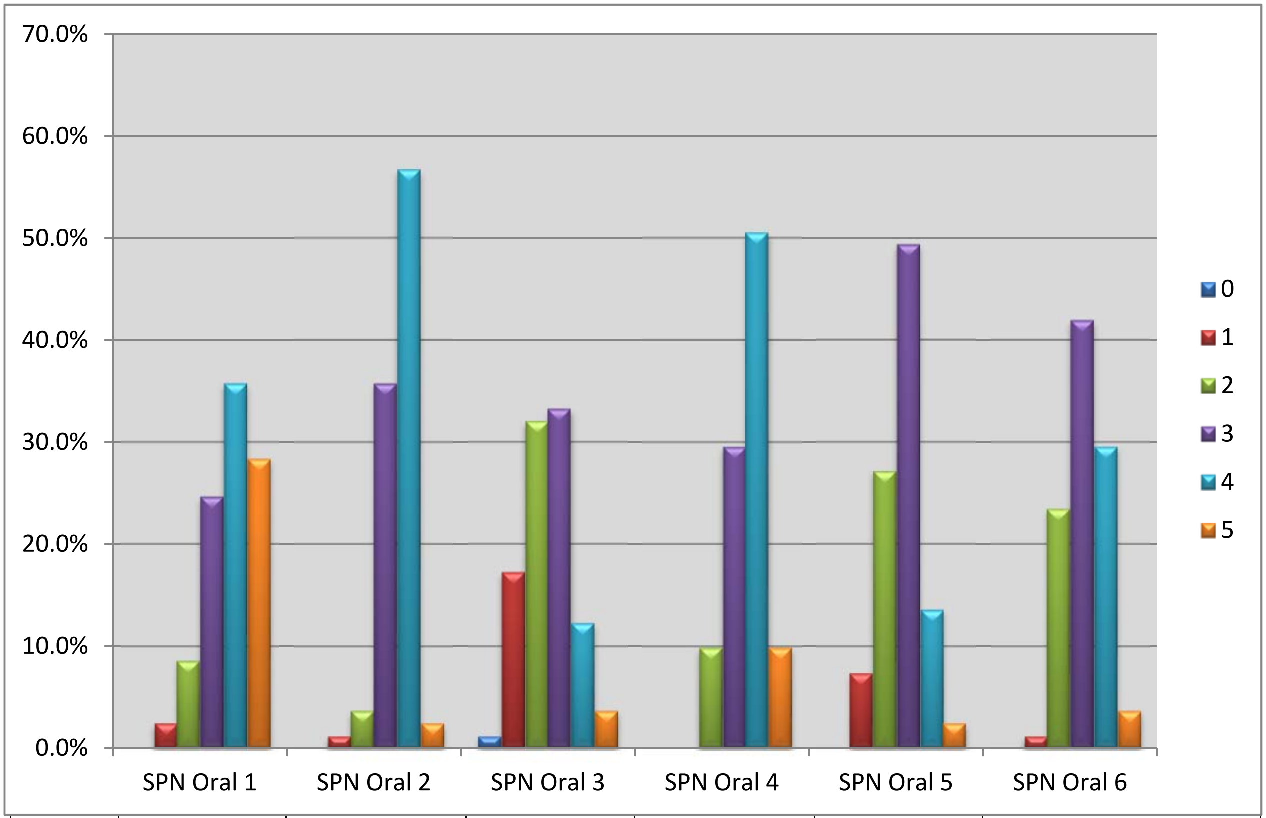
Spanish Speaking and Listening

The rubric for writing proficiency in Spanish consists of six dimensions, each scored on a six-point scale.

SUMMARY OF SCORES BY DIMENSION

Five faculty scorers scored 81 student oral interview exams from two courses from the Fall 2012 semester, SPN 102 and SPN 201. Sixty interviews (74.1%) were scored by multiple scorers. Figure 6.2 provides the score distributions for each dimension for oral interviews that were scored on that dimension (interviews scored as not applicable [NA] are not included).

SPANISH SPEAKING AND LISTENING RESULTS BY DIMENSION FOR APPLICABLE SCORES ONLY



	SPN Oral 1	SPN Oral 2	SPN Oral 3	SPN Oral 4	SPN Oral 5	SPN Oral 6
0	0.0%	0.0%	1.2%	0.0%	0.0%	0.0%
1	2.5%	1.2%	17.3%	0.0%	7.4%	1.2%
2	8.6%	3.7%	32.1%	9.9%	27.2%	23.5%
3	24.7%	35.8%	33.3%	29.6%	49.4%	42.0%
4	35.8%	56.8%	12.3%	50.6%	13.6%	29.6%
5	28.4%	2.5%	3.7%	9.9%	2.5%	3.7%
25th %tile	3	3	2	3	2	2.5
50th %tile	4	4	2	4	3	3
75th %tile	5	4	3	4	3	4
Mode	4	4	3	4	3	3
N	81	81	81	81	81	81

Figure 6.2 Distribution of Scores for Spanish Speaking and Listening, Applicable Scores Only

RESULTS BY DIMENSION

SPN Oral 1 Listening Comprehension

Scores on listening comprehension were the highest of the five dimensions. No students scored a zero on this dimension. Just fewer than three in 100 students scored a level one, meaning that their responses were always inappropriate. Fewer than one in 10 students scored a two, indicating that the responses provided during the interview were mostly not appropriate. One in four students scored a three on this dimension, indicating that responses were often inappropriate. Just over one-third of the students gave responses that were mostly appropriate (scores of four) and almost three in 10 of the interviews scored indicated that students almost always provided appropriate responses (scores of five).

SPN Oral 2 Pronunciation

Just as for dimension one, no students scored a zero on this dimension. One out of 100 students demonstrated utterances that were almost incomprehensible, receiving scores of one. One in 25 students demonstrated severe interference from their native language in the pronunciation of the Spanish words used in the interview (scores of two). Substantial interference from the native language was observed in just over one-third of the interviews (scores of three). Almost six out of 10 students received scores of four, demonstrating occasional interference from the native language. Showing little or no interference from the native language, one out of 40 students scored a five on this dimension.

SPN Oral 3 Vocabulary, Variety of items and expressions

One student scored a zero for this dimension. Nine students in 50 showed a severe lack of course-appropriate vocabulary (score of one). Almost one-third of the students received scores of two, insufficiently using course-appropriate vocabulary. One in three students exhibited occasional use of course-appropriate vocabulary and received scores of three. Three in 25 students' interviews were given scores of four, indicating that there was frequent use of course-appropriate vocabulary. Finally, one out of 25 students showed extensive use of course-appropriate vocabulary and received scores of five.

SPN Oral 4 Vocabulary, Proper Use

No students scored a zero or one for this dimension. One in 10 students received scores of two, using vocabulary with frequent errors. Three in 10 students exhibited more than occasional errors and received scores of three. Half of the students' interviews were given scores of four, indicating that there were few vocabulary errors. Finally, one out of 10 students showed almost no errors in their vocabulary use and received scores of five.

SPN Oral 5 Grammar

No students scored a zero on the Grammar dimension. Two in 27 exhibited an almost non-existent usage of correct grammar (scores of one). Almost three in 10 students had substantial grammatical errors that obscured the meaning of their statements (scores of two). Just under one-half of the students scored a level three, indicating that there were several errors and some avoidance of structures during the oral interview. One or two significant errors were observed in just under one in seven interviews (scores of four). Finally, two students had no significant errors (scores of five).

SPN Oral 6 Fluency

The same as for dimensions SPN Oral 1 through SPN Oral 5, this dimension had no students scoring at a level zero. One student in 100 was assigned a score of one for this dimension, with incomprehensible utterances. Almost one quarter of students exhibited many pauses with communication breakdown (scores of two). Just over four in 10 students were given scores of three, indicating that there was frequent hesitation, though no significant breakdown of communication, during their interviews. Demonstrating only slight hesitation with the natural pauses, three in 10 students received scores of four. One out of 25 students exhibited only natural pauses (scores of five).

Correlation between Dimensions

All dimension scores were correlated with each other at the .05 (SPN Oral 3 and SPN Oral 4) or .01 (all others) level of significance. The range of correlation coefficients was .281 to .725, with SPN Oral 1 Listening Comprehension and SPN Oral 6 Fluency having the highest correlation. See Appendix 6.B at the end of this chapter for a complete presentation of correlation coefficients.

Demographic and Preparedness Findings

There were no statistically significant differences between the means, medians, and the score distributions of honors vs. non-honors students. When comparing the scores of transfer vs. freshman-start students, there was a statistical difference between the scores of these groups on SPN Oral 1 Listening Comprehension with freshman-start students scoring higher on this dimension. There was also a statistical difference between the scores of one racial group and the others on SPN 5 Grammar with self-reported aliens scoring higher.

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 (21.6% of the sample), 31 – 60 credit hours (35.2% of the sample), 61 – 90 (33.0% of the sample), and over 90 credit hours (10.2% 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 for any of the Spanish speaking and listening rubric dimensions. Looking at Spearman rho correlation coefficients, the number of total hours completed was not significantly correlated with any of the rubric dimensions.

GPA was significantly correlated with SPN Oral 1 (.363**), SPN Oral 2 (.336*), SPN Oral 3 (.339*) and SPN Oral 5 (.350*). SAT-Verbal was significantly correlated with SPN Oral 1 (.460*) and SPN Oral 6 (.256*), and SAT-Math was significantly correlated with SPN Oral 1 (.375**), SPN Oral 2 (.257*), and SPN Oral 5 (.286*). ACT scores were significantly correlated with SPN Oral 1 (.571*), SPN Oral 2 (.624*), and SPN Oral 4 (.555*).

Interrater Reliability

In some instances, multiple reviewers scored the student interviews so that interrater reliability could be assessed (60 or 74.1%). Table 5.3 shows the reliability measures for Spanish Writing.

Table 6.3 Interrater Reliability for Spanish Speaking and Listening

Dimension	Percent Agreement	Plus Percent Adjacent	Krippendorff's Alpha
SPN Oral 1 Listening Comprehension	47.5%	84.7%	.638
SPN Oral 2 Pronunciation	27.1%	79.7%	.057
SPN Oral 3 Vocabulary, Variety of items and expressions	28.8%	62.7%	.434
SPN Oral 4 Vocabulary, Proper use	35.6%	74.6%	.174
SPN Oral 5 Grammar	25.4%	76.3%	.365
SPN Oral 6 Fluency	32.2%	74.6%	.365

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 Krippendorff's Alpha measure scorer agreement. The UNCW benchmarks are .67 for Krippendorff's alpha. See Appendix B of the General Education Assessment 2013 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, no dimension met this standard. Looking at percent agreement plus adjacent (that is, the scores that were within one level of each other), we find that no dimensions had 90% of scores within one level of each other.

Discussion

This was the first independent study using the Spanish Speaking and Listening Assessment Rubric. Table 5.4 shows the percent of work products scored at a level 3 or higher for each dimension. Level 3 is the benchmark for proficiency.

Table 6.4 Spanish Speaking and Listening Percent of Sample Scored at or above 3 and 4

Dimension	% of Work Products Scored 3 or higher	% of Work Products Scored 4 or higher
SPN Oral 1 Listening Comprehension	88.9%	64.2%
SPN Oral 2 Pronunciation	95.1%	59.3%
SPN Oral 3 Vocabulary, Variety of items and expressions	49.4%	16.0%
SPN 4 Oral, Vocabulary, Proper use	90.1%	60.5%
SPN Oral 5 Grammar	65.4%	16.1%
SPN Oral 6 Fluency	75.3%	33.3%

The results indicate that Listening Comprehension, Pronunciation, and Proper Use of Vocabulary are areas of relative strength, and that Vocabulary—Variety is an area of relative weakness. Interrater reliability statistics indicate reliability of scores can be improved, 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.

Appendix 6.A Second Language Rubrics

French Speaking Assessment

Listening Comprehension (appropriateness of responses to questions/statements)		**Pronunciation** (production of individual sounds, intonation, stress)		**Vocabulary** (breadth & variety of lexical items, idiomatic expressions, & use of French appropriate to level)		**Grammar** (Control of course- appropriate structures, forms, & syntax; spelling) Can you apply the grammar learned correctly?		**Fluency** (flow of speech, level of hesitation, & use of Spanish)	
5	Responses almost always appropriate	5	little or no interference from native language	5	extensive use of course-appropriate vocabulary, with almost no errors	5	one or two significant errors	5	only natural pauses
4	responses mostly appropriate	4	occasional interference from native language	4	frequent use of course-appropriate vocabulary, with few errors	4	some significant errors, but meaning clear	4	slight hesitation with natural pauses
3	responses often inappropriate	3	substantial interference from native language	3	occasional use of course-appropriate vocabulary, with more than occasional errors	3	several significant errors and/ or avoidance of structures	3	frequent hesitation; no significant breakdown of communication
2	responses mostly inappropriate	2	severe interference from native language	2	insufficient use of course-appropriate vocabulary, with frequent errors	2	substantial errors; meaning is obscured	2	many pauses with communication breakdown
1	responses always inappropriate	1	utterances are almost incomprehensible	1	lack of course-appropriate vocabulary, with substantial errors	1	correct usage of grammar almost non-existent	1	utterances are almost incomprehensible
0		0		0	extensive use of course-appropriate vocabulary, with almost no errors	0	correct usage of grammar non- existent	0	

Spanish Speaking Assessment

Listening Comprehension (appropriateness of responses to questions/statements)		**Pronunciation** (production of individual sounds, intonation, stress)		**Vocabulary**				**Grammar** (Control of course-appropriate structures, forms, & syntax; spelling) Can you apply the grammar learned correctly?		**Fluency** (flow of speech, level of hesitation, & use of Spanish)	
				(Variety of course-appropriate lexical items and idiomatic expressions) Are you using abundant vocabulary learned in your class?		(Proper use) Are you staying away from false cognates or words that don't exist and using words correctly?					
5	Responses almost always appropriate	5	little or no interference from native language	5	extensive use	5	almost no errors	5	one or two significant errors	5	only natural pauses
4	responses mostly appropriate	4	occasional interference from native language	4	frequent use	4	few errors	4	some significant errors, but meaning clear	4	slight hesitation with natural pauses
3	responses often inappropriate	3	substantial interference from native language	3	occasional use	3	more than occasional errors	3	several significant errors and/ or avoidance of structures	3	frequent hesitation; no significant breakdown of communication
2	responses mostly inappropriate	2	severe interference from native language	2	insufficient use	2	frequent errors	2	substantial errors; meaning is obscured	2	many pauses with communication breakdown
1	responses always inappropriate	1	utterances are almost incomprehensible	1	severe lack of use	1	substantial errors	1	correct usage of grammar almost non-existent	1	utterances are almost incomprehensible
0		0		0	complete lack of use	0	correct usage of vocabulary non-existent	0	correct usage of grammar non-existent	0	

APPENDIX 6.B CORRELATIONS BETWEEN RUBRIC DIMENSIONS

Spearman rho Rank Order Correlation Coefficients for FRH Speaking and Listening

Correlations

			FRH1	FRH2	FRH3	FRH4	FRH5
Spearman's rho	FRH1	Correlation Coefficient		.573**	.686**	.690**	.587**
		N		25	25	25	25
	FRH2	Correlation Coefficient	.573**		.632**	.614**	.614**
		N	25		25	25	25
	FRH3	Correlation Coefficient	.686**	.632**		.767**	.737**
		N	25	25		25	25
	FRH4	Correlation Coefficient	.690**	.614**	.767**		.605**
		N	25	25	25		25
	FRH5	Correlation Coefficient	.587**	.614**	.737**	.605**	
		N	25	25	25	25	

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

Spearman rho Rank Order Correlation Coefficients for SPN Speaking and Listening

Correlations

			SP1	SP2	SP3	SP4	SP5	SP6
Spearman's rho	SP1	Correlation Coefficient		.618**	.395**	.583**	.647**	.725**
		N		81	81	81	81	81
	SP2	Correlation Coefficient	.618**		.386**	.558**	.530**	.415**
		N	81		81	81	81	81
	SP3	Correlation Coefficient	.395**	.386**		.281*	.621**	.486**
		N	81	81		81	81	81
	SP4	Correlation Coefficient	.583**	.558**	.281*		.511**	.406**
		N	81	81	81		81	81
	SP5	Correlation Coefficient	.647**	.530**	.621**	.511**		.611**
		N	81	81	81	81		81
	SP6	Correlation Coefficient	.725**	.415**	.486**	.406**	.611**	
		N	81	81	81	81	81	

**. 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, and new recommendations.

UNCW Student Abilities on Learning Goals

Five of the eight UNCW Learning Goals were assessed in 2012-2013 using course-embedded assignments. Table 7.1 combines the results from all studies, and presents the percentage 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/ Benchmark	Dimension	% of Work Products Scored at Proficiency Level or Higher
PSY 105 *(Benchmark = 2)*	IL1 Determine Extent of Information Needed	78.3%
	IL2 Access Needed Information	75.6%
	IL3 Evaluate Information and Sources	75.1%
	IL4 Use Information Effectively	78.4%
	IL5 Access and Use Information Ethically	75.6%
SED 372 *(Benchmark = 3)*	IL1 Determine Extent of Information Needed	40.6%
	IL2 Access Needed Information	6.3%
	IL3 Evaluate Information and Sources	18.8%
	IL4 Use Information Effectively	43.7%
	IL5 Access and Use Information Ethically	40.6%
ANTL 207, COM 160, ECN 222, ENG 230, FST 110, PSY 105, THR 121 *(Benchmark = 2)*	CT 1 Explanation of Issues	57.2%
	CT2a Evidence: Selecting and Using	53.8%
	CT2b Evidence: Critically Examining for Viewpoint	34.8%
	CT3a Influence of Assumptions	32.0%
	CT3b Influence of Context	35.4%
	CT4 Student’s Position	39.0%
	CT5 Conclusions and Related Outcomes	35.4%
FST 110, MUS115, THR 121 *(Benchmark = 2)*	WC1 Context of and Purpose for Writing	76.4%
	WC2 Content Development	75.5%
	WC3 Genre and Disciplinary Conventions	72.7%
	WC4 Sources and Evidence	59.4%
	WC5 Control of Syntax and Mechanics	83.9%
ACG 445, NSG 415, SED 372 *(Benchmark = 3)*	WC1 Context of and Purpose for Writing	44.4%
	WC2 Content Development	32.1%
	WC3 Genre and Disciplinary Conventions	30.9%
	WC4 Sources and Evidence	43.2%
	WC5 Control of Syntax and Mechanics	46.9%
FRH 201 *(Benchmark = 3)*	FRH Oral 1 Listening Comprehension	100.0%
	FRH Oral 2 Pronunciation	99.0%
	FRH Oral 3 Vocabulary	96.0%
	FRH Oral 4 Grammar	88.0%
	FRH Oral 5 Fluency	100.0%
SPN 102, SPN 201 *(Benchmark = 3)*	SPN Oral 1 Listening Comprehension	97.5%
	SPN Oral 2 Pronunciation	98.8%
	SPN Oral 3 Vocabulary, Variety of items and expressions	81.4%
	SPN 4 Oral, Vocabulary, Proper use	100%
	SPN Oral 5 Grammar	92.7%
	SPN Oral 6 Fluency	98.8%

For 12 dimensions (41.4%), 75% or more of the student work met the benchmark, and for 21 (72.4%), 50% or more of the student work met the benchmark. The highest levels of work meeting the benchmarks were in the Second Language study: greater than 90% of the work products were scored at or above the benchmark for nine of the 11 dimensions. Comparing the results within each rubric indicates specific areas that need additional coverage and opportunities for practice.

The Critical Thinking dimensions in particular stand out with lower percentages of work meeting the benchmarks: five of the seven dimensions had less than 40% of work products meeting the benchmark. The percentages of work meeting the benchmark for these dimensions ranges from 32.0% for CT3a Influence of Assumptions at the low end, to 57.2% for CT1 Explanation of Issues at the high end. Only CT1 and CT2a were cited by the scorers as fitting well with the assignments, without requiring additional scoring guidelines. CT2b and CT4 were not scored for one assignment and CT5 was not scored for another, but different, assignment.

Though the courses from which Critical Thinking evidence was drawn were all at the 100- and 200-level, the sample included students from all four class levels. Evidence from this study does not show that critical thinking skills significantly increased with credit hours completed. There were no significant correlations between the Critical Thinking rubric dimension scores and the total number of hours completed by students, though three dimensions were negatively correlated: CT1, CT2b, and CT4. This suggests that student critical thinking skills do not significantly improve over the course of their UNCW career, and suggests that students need opportunities to practice the important critical thinking skills of critically examining evidence and positions for viewpoint, context, and assumptions; exploring both one’s own perspective and those of others; and logically relating the conclusions drawn to the evidence and perspectives discussed.

Scorer Feedback on Process

Table 7.2 provides combined results for the survey items for the Information Literacy, Critical Thinking, Thoughtful Expression (Written), and Second Language scoring processes.

Table 7.2 Scorer Feedback on General Education Assessment Process

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The invitation to volunteer accurately described the experience.	0 (0%)	0 (0%)	0 (0%)	12 (75%)	4 (25%)
The timing of the invitation gave adequate opportunity to arrange for attending workshops and scoring.	0 (0%)	0 (0%)	0 (0%)	12 (70.6%)	5 (29.4%)
The norming session adequately prepared me for what was expected of me during the scoring session.	0 (0%)	0 (0%)	1 (5.9%)	4 (23.5%)	12 (70.6%)
The scoring session was well-organized.	0 (0%)	0 (0%)	0 (0%)	12 (70.6%)	5 (29.4%)
The structure of the scoring made it reasonable to work for the full time.	0 (0%)	0 (0%)	1 (5.9%)	7 (41.2%)	9 (52.9%)
When I had questions, one of the leaders was available to answer it.	0 (0%)	0 (0%)	0 (0%)	2 (12.5%)	14 (87.5%)
When I had questions, the question was answered.	0 (0%)	0 (0%)	0 (0%)	2 (12.5%)	14 (87.5%)
I was comfortable scoring student work products from outside my discipline on the board Learning Goals.	0 (0%)	1 (5.9%)	0 (0%)	7 (41.2%)	9 (52.9%)
The process is an appropriate way to assess students on the UNCW Learning Goals.	0 (0%)	0 (0%)	1 (5.9%)	10 (58.8%)	6 (35.3%)
This process is valuable in improving student learning.	0 (0%)	0 (0%)	1 (6.25%)	11 (68.75%)	4 (25%)
I would participate in this process again.	0 (0%)	0 (0%)	0 (0%)	4 (23.5%)	13 (76.5%)
I would recommend participating in this process to my colleagues.	0 (0%)	0 (0%)	1 (5.9%)	4 (23.5%)	12 (70.6%)

There were also two open-ended questions on the survey and a section for comments or suggestions. The results of the Likert-scale, open-ended questions, and comments or suggestions are discussed below.

Scorers reported high levels of agreement regarding most aspects of the norming and scoring processes. Regarding the invitation to volunteers, 12 scorers (75%) agreed and four strongly agreed that the invitation to volunteer accurately described the experience they had during the workshop One participant chose not to answer. Additionaly, 70.6% (n=12) agreed and five participants strongly agreed that the timing of the invitation gave adequate opportunity to arrange for attending workshops and scoring.

On the aspect of the norming session 12 scorers or 70.6% strongly agreed and four participants agreed that it adequately prepared them for the scoring session expectations while one participant was neutral on the subject. When it came to organization all scorers agreed or strongly agreed that the scoring session was well organized with 70.6% (n=12) agreeing and 20.4% (n=5) strongly agreeing. Regarding session structure and time, 52.9% (n=9) of scorers strongly agreed that the structure of the scoring made it reasonable to work for the full time while seven scorers, 41.2%, agreed and one scorer responded with neutral. All scorers either agreed, 87.5% (n=14), or strongly agreed, 12.5% (n=2), that when they had questions the question was answered and that a leader was available to answer it.

Regarding UNCW Learning Goals all but one scorer either strongly agreed (52.9%) or agreed (41.2%) that they were comfortable scoring student work products outside of their discipline on the board of Learning Goals. Also 58.8% of scorers (n=10) agreed and 35.3% (n=6) strongly agreed that the process was an appropriate way to assess students on the UNCW Learning Goals. One participant responded neutral to this statement. When asked if this process was valuable in improving student learning 68.75% (n=11) agreed, 25% (n=4) strongly agreed, and 6.25% (n=1) was neutral.

Regarding continued participation and recommendation all scorers agreed or strongly agreed that they would participate in the process again with 76.5% (n=13) strongly agreeing and 23.5% (n=4) agreeing. Also a majority of scorers, 70.6% (n=12), strongly agreed, 23.5% (n=4) agreed, and one scorer responded with neutral that they would recommend participating in this process to their colleagues. When asked what parts of the process worked the best there were three major themes. Seven scorers felt the norming process worked the best, seven more felt that either working with a partner or team collaboration on student artifacts was the best working process, and four other scorers felt that the scoring workshop in a whole was the best part. Other well-noted processes were the organization of the workshop, the amount of time allowed to complete work, discussions on the artifacts, and receiving feedback. The last open-ended question asked in what ways the scoring process could be improved. Two scorers commented that the scoring process could be improved by looking at what anchor or model papers were used and how they support the training process. Three other scorers commented on improvement of the rubric and its clarity and two more scorers commented that they felt rushed on the timing of scoring of one of their packets. Other improvements suggested consisted of adding shorter papers (less than 15 pages) to the packets, having professors align their assignments with the assignment rubric beforehand, altering the starting time of the workshop, and the room temperature was too cold and it was noisy making it distracting for some scorers. Overall, participants described the scoring event's invitation, expectations, structure, and applicability positively.

INTERRATER RELIABILITY

Table 7.3 combines the interrater reliability findings from all 2012-2013 studies, arranged in descending order by Krippendorff's alpha.

Table 7.3 Interrater Reliability

Dimension	Krippendorff's Alpha	Percent Agreement	Plus Percent Adjacent
FRH Oral 2 Pronunciation	0.947	90.0%	100%
FRH Oral 5 Fluency	0.935	80.0%	100%
FRH Oral 1 Listening Comprehension	0.866	80.0%	100%
IL2 Access Needed Information	0.771	55.6%	100%
WC4 Sources and Evidence	0.761	56.8%	97.3%
WC2 Content Development	0.746	61.9%	100%
FRH Oral 4 Grammar	0.744	50.0%	100%
IL4 Use Information Effectively	0.722	66.7%	100%
IL5 Access and Use Information Ethically	0.674	55.6%	94.4%
FRH Oral 3 Vocabulary	0.670	30.0%	90.0%
SPN Oral 1 Listening Comprehension	0.638	47.5%	84.7%
WC3 Genre and Disciplinary Conventions	0.559	50.0%	95.2%
WC1 Context of and Purpose for Writing	0.530	57.1%	92.9%
CT3a Influence of Assumptions	0.524	46.6%	86.2%
IL1 Determine Extent of Information Needed	0.521	55.6%	94.4%
WC5 Control of Syntax and Mechanics	0.519	57.1%	95.2%
CT3b Influence of Context	0.500	34.5%	89.7%
CT4 Student's Position	0.488	59.3%	92.6%
CT5 Conclusions and Related Outcomes	0.481	45.3%	94.3%
SPN Oral 3 Vocabulary, Variety of items and expressions	0.434	28.8%	62.7%
CT 1 Explanation of Issues	0.415	34.5%	91.4%
IL3 Evaluate Information and Sources	0.413	53.3%	93.3%
SPN Oral 5 Grammar	0.365	25.4%	76.3%
SPN Oral 6 Fluency	0.365	32.2%	74.6%
CT2a Evidence: Selecting and Using	0.342	32.8%	81.0%
CT2b Evidence: Critically Examining for Viewpoint	0.335	30.0%	86.8%
SPN 4 Oral, Vocabulary, Proper use	0.174	35.6%	74.6%
SPN Oral 2 Pronunciation	0.057	27.1%	79.7%

The highest IRR results were for the French Speaking and Listening rubric, with all five dimensions met the Krippendorff's agreement benchmark. On the Information Literacy rubric, three of the five dimensions met the Krippendorff's benchmark. On the Written Communcation rubric, two of the five dimensions met the Krippendorff's benchmark. The higher Krippendorff's alpha scores indicate a greater chance that the scores from different scorers were in agreement not because of chance.

The percentage of scores for French Oral, Information Literacy, and Written Communication that were in agreement or within one score level was above 90%, which indicates good interrater reliability. IRR was good for Critical Thinking was good as well, with all dimensions having greater than 80% of scores either in agreement or within one score level of each other. The IRR for Spanish Oral was the lowest, but it bears mention that, for some of the Spanish interviews, there were as many as four scorers, making IRR more difficult to achieve than for the usual pair of scorers.

Follow up on Previous Recommendations

The following explains the progress made on the recommendations from last year.

2012 Recommendations

- *The General Education Assessment office will disaggregate the Inquiry data for Dimension 6 to analyze possible differences between courses and/or sections.*

For this dimension, IN6, there was a significant difference between the scores from the two courses sampled, with work products from Biology courses scoring higher. This is could be explained by the type of assignment collected from each course; the Biology assignments were out-of-class lab reports and the Chemistry assignments were in-class lab practicals. Likely, the extended time and instructions that the Biology students received played some role in the higher scores on those work products.

- *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 text for this email was drafted. Due to restrictions on how emails are sent to all faculty, the email has not yet been sent.

- *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.*

During the 2012-2013 academic year, the Spanish and French departments adopted the same rubric for scoring the speaking and listening of students during course oral interviews. The results from the first cross-departmental use of these rubrics are reported in this document.

- *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.*

These reports were completed.

- *A Director of University Studies position should be created and filled by July 1, 2013.*

The filling of this position was postponed.

NEW RECOMMENDATIONS

Based on the discussion conclusions that (1) more instructions were needed in assignments to elicit higher performance in WC2 Content Development, WC3 Genre and Disciplinary Conventions, and WC4 Sources and Evidence and (2) assignments including directions such as "critically analyze," "include rationale," "summarize," "evaluate," "critique," and "elaborate" produced higher-scoring work, the following recommendation was adopted by the University Studies Advisory Board:

The results and analysis of the assessment process must be disseminated more purposefully and broadly so that faculty members can address these findings in their courses. A team consisting of the Associate Vice Chancellor and Dean of Undergraduate Studies, the Director of General Education Assessment, the Chair of the University Studies Advisory Committee, and the Undergraduate Studies Liaison to University Studies will present findings and suggestions at each department/school faculty meeting during the 2014-2015 academic year. Cumulative results from 2011-2013 for Written Communication were chosen for the first round of presentations, as they are relevant to all courses.

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.