Measuring Writing as a Representation of Disciplinary Knowledge

Kathleen Burke[1], Judith Ouellette[1], Wendy Miller[1], Cy Leise[2], and Tris Utschig[3]

Abstract

Most students are exposed to different writing genres as well as to grammar and other writing-specific skills during their required composition course(s). Discipline-specific writing skills are often not taught, rather they are assumed to be prerequisite knowledge by disciplinary instructors. Rubrics are often used by instructors to measure these skills; however, few rubrics integrate measurement of both general and discipline-specific skills. One such rubric has been designed by the Academy of Process Educators. The present research examines the empirical reliability and validity of this rubric. It was hypothesized that consistency of ratings would be achieved for three different types of writing by four raters from different disciplines when read in a counterbalanced pattern. The Written Communication Value Rubric *from the Association of American Colleges and Universities (AAC&U, 2012) was also utilized to make possible the evaluation of concurrent validity of the two rubrics. Different statistical methods were utilized to analyze the ratings. The research methodology utilized provides a model for empirically exploring the reliability and validity of other rubrics. Overall, our findings indicate that individually raters were consistent on both rubrics, but the inter-rater reliability was not strong. The ratings were less dispersed on the Academy rubric than on the AAC&U rubric but "concurrent" validity, that is significant positive correlations between the rubrics, were clearly present. Some of the implications of the study include the importance of encouraging careful rater and inter-rater training, using counter-balanced sequences to avoid bias from repetition, ensuring that writing assignments reflect clear performance criteria in regards to rubric performance areas, and assuring that writing pre-requisites are carefully integrated throughout the program's curriculum and assessment strategy.*

Overview

Educators often measure learners' mastery of disciplinary knowledge through writing. In discipline-based writing assignments, learners are expected to integrate both writing skills and disciplinary knowledge. These skills and knowledge are often assessed and evaluated using rubrics either designed by instructors or downloaded from public websites. An important concern is that many of these measures lack empirical support for reliability and validity, which reduces the value of data collected with these measures. Conversely, well-constructed and empirically tested rubrics are a logical and efficient method for confidently assessing and evaluating the writing skills of learners across disciplines for many kinds of written assignments used in college and graduate school programs.

The purpose of this study is to examine the reliability of a new version of the Academy of Process Educators Writing Rubric originally presented by Burke and Nancarrow (2007). The new rubric was designed to measure both fundamental writing skills as well as depth of disciplinary knowledge. The current research seeks to determine whether the design of this rubric allows for measurement by raters across different disciplines as well as for different writing purposes. That is, will the use of this rubric yield consistent measurement when utilized for different types of writing? Moreover, will its use produce inter-rater reliability with raters from different disciplines?

To answer these questions, we first explore the purpose of writing in college and graduate school and the obstacles that students often face when trying to write well. We then discuss how writing performance is usually measured and the development of the rubrics that are analyzed in this paper. Next we present the methods we used, our findings on the reliability of rubrics, the agreement between raters, and comparisons of the raters' consistency within and across rubrics. Finally we discuss the implications of our findings and next steps.

Background

Purposes of Writing in College and Graduate School

Required composition courses are usually the only situations that explicitly expose students to writing genres as well as to grammar and other writing-specific skills. These courses are designed to prepare learners to apply their written communication skills. In most college and graduate courses, however, the assessment of writing as a skill is secondary to the representation and organization of disciplinary knowledge. Learners must integrate their accuracy of knowledge, incisiveness of analysis, quality of achievements, and clarity of proposals to produce quality writing in a disciplinary context. This integration, however, is not often taught by instructors; it is assumed to be prerequisite knowledge.

[1] State University of New York at Cortland
[2] Bellevue University
[3] Georgia Institute of Technology

Graham and Harris (2000a) provide an overview of the many benefits of the flexible medium of writing in their introduction to a special issue on writing development published in the *Educational Psychologist.* In this special issue, McCutchen (2000) argues, on the basis of a literature review, that novice writers lack fluency in language skills. Moreover, they do not have the extensive knowledge base in long-term memory of more expert writers. Apple, Beyerlein, Leise, and Baehr (2007) present a model of learning skills in which language development is a foundation for skill development across all learning domains. Consistent with McCutchen's argument based on memory resources, the language development problems of many learners cause difficulty for them when they move to higher-level skills unless this transition is facilitated in a developmentally logical manner. A relatively universal writing rubric must address advanced skills, including meta-skills such as being able to select and use the appropriate genre when faced with novel challenges. McCutchen illustrates how expert writers set "constraining" criteria during their pre-writing analysis phase. Less experienced writers and those with less long-term knowledge are unaware of or do not consistently use this strategy.

Graham and Harris (2000b) discuss a methodology that is thought to guide self-regulation of writing. Three processes must be managed well via predictable steps for this to occur. The first process is environmental. The writer must choose to use either paper or computer and to work in a group or alone. The second process is behavioral. The writer must manage the motor acts involved in writing or typing. The third process is personal. The writer needs to manage his or her cognitive beliefs and affective states associated with writing that may influence motivation.

Graham and Harris also review evidence that "transcription skills"—spelling, grammar, etc.—can be barriers to the growth of writing skills. When these barriers are due to learning disabilities these authors suggest that new technologies such as speech recognition software may help some learners bridge these basic skill issues even if they cannot be fully resolved.

Bruning and Horn (2000) explore the problems related to writing motivation and offer four clusters of conditions: (1) nurturing functional beliefs about writing, (2) fostering engagement by using authentic writing tasks, (3) providing a supportive context for writing, and (4) creating a positive emotional environment. These authors argue that many learners have experienced writing as an anxiety-arousing activity because they are not taught how to self-manage in ways that will lead to intrinsic motivation—a belief that they can improve and succeed despite unclear expectations from poorly-designed assignments and other challenges. The premises of these authors are consistent with the skill facilitation and assessment recommendations of process educators, e.g., Smith and Apple (2007) regarding quality learning environment design, Burke (2007) on getting student buy-in, and Hanson and Moog (2007) on guided inquiry.

Measurement Quality Issues and Writing Rubrics

The instrument that is often used to measure writing is a rubric. The context in which rubrics are applied can influence rater reliability such as when many documents of the same type must be reviewed in a brief time. Furthermore, the quality of a rubric may be clear for certain purposes but not for others, e.g., different genres or disciplines. To provide writers with consistent feedback the instrument that is used to evaluate their writing must be able to be used for multiple types of writing. Moreover, the rubric needs to be constructed purposefully. Using a carefully constructed rubric, Cho, Schunn, and Wilson (2006) demonstrated that multiple peer evaluations approximated instructor ratings. Finally, the rubric needs to provide feedback to the writer regarding their performance with respect to both disciplinary writing criteria as well as their level of basic writing skills.

Burke and Nancarrow (2007) presented a model of quality writing, a holistic rubric, and an analytic rubric for writing performance in discipline contexts. The rubric was developed by a diverse group of faculty from multiple disciplines. The rubrics presented by the authors measure both the mechanics of writing as well as disciplinary content. After the Academy of Process Educators used the analytic rubric presented by Burke and Nancarrow to assess multiple professional papers, Academy members agreed that a condensed version of the rubric would be of more value to instructors. This new version of the Academy rubric was discussed during the 2010 Process Education Conference at a session regarding inter-rater reliability of the revised Academy rubric. The feedback from this session has prompted our current research examining the inter-rater reliability of the rubric as well as comparing the reliability of the Academy rubric to a publicly used rubric, the American Association of Colleges and Universities *Written Communication Value Rubric* (AAC&U Rubric). Both the Academy rubric and the AAC&U rubric are included in the Appendix.

Methods

Three different student writing types (book reviews, essays, and research papers) were rated using two different rubrics (the Academy vs. AAC&U). Four raters, two females and two males, from three different institutions and representing four different areas of study (economics,

engineering, geography, and psychology and human services) independently rated samples of each writing type using each rubric.

Writing Rubrics

The Academy Writing Rubric consists of five Performance Areas: (1) Audience-Orientation, (2) Discipline Knowledge, (3) Analytical Quality/Critical Thinking, (4) Synthetic Quality, and (5) Use of Language. Each Performance Area contains three sub-categories. For example, Audience Orientation is comprised of Thesis Relevance, Thesis Clarity, and Cohesiveness of Perspective. The papers are rated on each sub-category using a four-point Likert scale where a one-word descriptor represents each number. For example, in the Thesis Relevance sub-category, a score of 1 is Marginal, 2 is Adequate, 3 is Valuable, and 4 is Visionary. The specific words assigned to each subcategory score can be seen on the Academy Rubric in the Appendix.

The *Written Communication Value Rubric* from the AAC&U (AAC&U, 2012) introduced in the previous section, is available on the Internet and was incorporated into this study to make possible the evaluation of concurrent validity of the two tools. It consists of five Performance Areas: (1) Content and Purpose for Writing, (2) Content Development, (3) Genre and Discipline Conventions, (4) Sources and Evidence, and (5) Control of Syntax and Mechanics. Every paper can be assigned a score of 1 (Benchmark) to 4 (Capstone) for each of the Performance Areas. Descriptors are provided for each cell within the rubric.

As we will discuss later, the five Performance Areas from the Academy and AAC&U rubrics will be paired based upon what is assessed. The following five Academy/AAC&U pairings will be examined: Audience-Orientation/Context of and Purpose for Writing, Discipline Knowledge/ Sources and Evidence, Analytical Quality and Critical Thinking/Genre and Disciplinary Conventions, Synthetic Quality/Content Development, Use of Language/Control of Syntax and Mechanics.

Writing Samples and Raters

Three types of writing were assessed using the two rubrics. The student work was gathered from course assignments at two different institutions in three different disciplines (psychology, economics, and psychology and human services). Students from a graduate counseling course were given a lengthy description of the book review assignment that included a recommended outline, and advice to focus on setting criteria by which to review the book and two rubrics. Burke and Nancarrow (2007) presented a model of quality writing, a holistic rubric, and an analytic rubric, all of which can be used to measure writing performance in disciplinary contexts. The book reviews ranged in length from four to ten pages. The essays were obtained from an undergraduate 400-level on-line health psychology course. The essays were part of a larger assignment which asked students to read a research article and a summary paragraph written by the instructor, and to respond to a question based on those readings. Essays ranged in length from one paragraph to two pages and the students were not provided with a specific rubric that would be used for assessment. The research papers came from an undergraduate 400-level strategic management course within the economics department. Students were given a one-page assignment description and were told that their work would be assessed using a departmental rubric. The papers ranged in length from ten to twenty-five pages.

Design

The research design required that all raters read 36 of each type of paper in a preset order, assessing them using both the Academy and the AAC&U rubrics at different times. The sample size was set at 36 to meet several research criteria: (1) statistical effects are reasonably stable with n more than 30, (2) raters could comfortably manage the total of 216 ratings (3 writing types x 2 rubrics x 36 sample documents), and (3) counterbalancing in combination with Latin Square sequences could be used to reduce bias due to carryover from repeated reading of each paper.

Each rater read the paper types (essay, book review, research paper) in counterbalanced sets of nine (ABC, BCA, CAB) changing to the alternate rubric at the end of each nine paper set. These patterns were filled in using randomized numbers from 1 to 36 representing the individual papers within each writing type to complete the rating sequence. The final sequence had the rater read 216 papers (36 papers × 3 types × 2 rubrics). This process was repeated until four distinct rating sequences were created. An example of a rating sequence is included in the Appendix. The rater began in Column 1, proceeding down the column, and ending at the bottom of Column 6. The papers were accessed from a central website and could be viewed as Microsoft Word documents or in .pdf form. The raters used an on-line version of the Academy rubric to assess each paper. The raters completed a Microsoft Excel spreadsheet with their scores from the AAC&U rubric.

Formal training on the use of the rubrics through a deliberate calibration activity or discussion among the raters purposefully did **not** occur. We wanted to determine the baseline agreement between the raters without introducing training.

Analysis and Results

Reliability of the Rubrics

Cronbach's alpha was used to determine internal consistency within the Academy rubric. That is, to determine whether the three sub-categories in each Performance Area of the Academy rubric could be collapsed into one measure for that Performance Area. Cronbach's alpha is defined as

$$\alpha = \frac{N \cdot \bar{c}}{\bar{v} + (N-1) \cdot \bar{c}}$$

where N is the number of items, $\bar{c}$ is the average inter-item covariance among the items and $\bar{v}$ equals the average variance.

A high value of alpha, 0.80 or greater, indicates that the individual items are reliable and can be collapsed into a single measure for the performance area. The results for Cronbach's alpha, presented in Table 1, indicate that the three subcategories within each Performance Area in the Academy rubric were reliable. Thus, the sub-categories were combined by averaging each rater's scores for a paper, resulting in one rating for each Performance Area.

The Cronbach's alpha measure was further used to determine the internal consistency within each rubric; that is, are the Performance Areas within each rubric highly related to each other indicating that they are measuring similar underlying constructs? The alpha values of 0.94 for the Academy rubric and 0.89 for the AAC&U rubric both indicate that there is internal consistency for each rubric. Interestingly, the Academy rubric has higher reliability in

Table 1 Results from Cronbach's Alpha Test on the Academy Rubric

Performance Area	Sub-Categories	Cronbach Alpha
Audience-Orientation	Thesis relevance	α = 0.83
	Thesis clarity	
	Cohesiveness of perspective	
Discipline Knowledge	Selection of citations	α = 0.85
	Depth of disciplinary knowledge	
	Representation of knowledge	
Analytical Quality/ Critical Thinking	Logic of development	α = 0.84
	Validity of evidence	
	Application of knowledge	
Synthetic Quality	Interpretation of evidence	α = 0.87
	Connection to thesis	
	Quality of insights/ conclusions	
Use of Language	Grammar/mechanics	α = 0.82
	Use of structure	
	Rhetorical eloquence	

Table 2 Standard Deviation of Raters' Scores

Academy Rubric			
Performance Area	**Research Paper**	**Book Review**	**Essay**
Audience-Orientation	0.6041	0.6375	0.7913
Discipline Knowledge	0.5914	0.7320	0.7400
Analytical Quality/Critical Thinking	0.5348	0.6526	0.5932
Synthetic Quality	0.6295	0.8300	0.8174
Use of Language	0.5949	0.6811	0.7348
AAC&U Rubric			
Performance Area	**Research Paper**	**Book Review**	**Essay**
Context of and Purpose for Writing	0.9306	0.9199	0.9996
Sources and Evidence	0.8499	0.8094	0.9108
Genre and Disciplinary Conventions	1.0215	0.9423	1.0151
Content Development	0.9554	1.0365	0.9574
Control of Syntax and Mechanics	0.6240	0.7612	0.8542

our sample than the nationally recognized AAC&U rubric. Three of the four raters had previous experience with the Academy rubric to evaluate journal submissions, however the raters received no specific training on the rubric for use in this study.

Agreement between Raters

We analyzed the agreement between the raters in three ways. First, to determine how close the rater scores were to each other, the standard deviation for each Performance Area was calculated and is presented in Table 2. Across all Performance Areas, the raters' scores were less dispersed when using the Academy rubric than when the AAC&U rubric was used.

Second, to determine the level of correlation among the raters' scores, we calculated Pearson correlation coefficients between the raters. Pearson's correlation coefficient, r, is calculated as

$$r = \frac{\sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n}}{\sqrt{\left(\sum X^2 - \frac{\left(\sum X\right)^2}{n}\right)\left(\sum Y^2 - \frac{\left(\sum Y\right)^2}{n}\right)}}$$

where r is a value between 0 and 1. The closer r is to 1, the higher the correlation, indicating a stronger relationship between the pairings. The correlations vary considerably, but generally have a stronger relationship on the Academy rubric (See Table 3) than on the AAC&U rubric (See Table 4).

Finally, to further measure the level of agreement among the four raters, Cohen's Kappa statistic (Cohen, 1960) was used. The Kappa statistic takes into account that agreement between raters can occur due to chance alone. The value of Kappa is defined as

$$K = \frac{p_o - p_e}{1 - p_e}$$

where P_o is the level of observed agreement and P_e is the level of agreement if the ratings were statistically independent; that is, they occurred by chance, for example, if the ratings were decided by a flip of a coin. Thus, the numerator represents the agreement between the raters when the probability that the agreement occurred by chance is removed. The denominator represents the possible level of agreement when chance is removed. Therefore, the closer the Kappa statistic is to 1, the higher the level of agreement between the raters. There is much debate in the literature about the validity of using the Kappa statistic for more than two raters (See Conger, 1980 and Fleiss, 1971).

Table 5 Percent Agreement for Rater Combinations across Performance Areas for the AAC&U and Academy Rubrics

AAC&U Rubric	Research Paper	Book Review	Essay
Context of and Purpose for Writing	0 of 6	0 of 6	0 of 6
Sources and Evidence	2 of 6 (44%* and 50%*)	2 of 6 (56%** and 50%**)	2 of 6 (61%*** and 39%*)
Genre and Disciplinary Conventions	0 of 6	1 of 6 (47%***)	1 of 6 (61%***)
Content Development	0 of 6	0 of 6	2 of 6 (39%** and 42%*)
Control of Syntax and Mechanics	2 of 6 (72%** and 50%**)	3 of 6 (56%**, 50%**, and 53%**)	0 of 6

Academy Rubric	Research Paper	Book Review	Essay
Audience-Orientation	0 of 6	2 of 6 (56%*** and 50%***)	0 of 6
Discipline Knowledge	1 of 6 (56%*)	2 of 6 (47%* and 72%***)	3 of 6 (42%**, 50%**, and 58%***)
Analytical Quality/ Critical Thinking	1 of 6 (77%***)	2 of 6 (39%** and 67%**)	1 of 6 (67%**)
Synthetic Quality	1 of 6 (64%*)	2 of 6 (44%** and 53%***)	2 of 6 (58%*** and 42%**)
Use of Language	1 of 6 (69%**)	1 of 6 (64%***)	4 of 6 (47%*, 58%**, 44%** and 42%**)

Note: *significant at the 10% level, **significant at the 5% level, ***significant at the 1% level

Table 3 Rater Correlation Using Academy Rubric

	Book Review				Essay				Research Paper			
Audience-Orientation	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**
Rater 1	1				1				1			
Rater 2	0.3641	1			0.4464	1			0.0414	1		
Rater 3	0.2387	0.6372	1		0.2469	0.2178	1		-0.0767	0.0327	1	
Rater 4	0.1025	0.4934	0.3938	1	0.3459	0.4095	0.1184	1	-0.2412	0.1629	0.2	1
Discipline Knowledge	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**
Rater 1	1				1				1			
Rater 2	0.3614	1			0.5823	1			-0.0565	1		
Rater 3	0.3776	0.6668	1		0.4992	0.4336	1		0.2394	-0.1145	1	
Rater 4	0.4461	0.6053	0.6439	1	0.5581	0.3547	0.314	1	0.2399	0.2759	0.0303	1
Analytical Quality/ Critical Thinking	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**
Rater 1	1				1				1			
Rater 2	0.1294	1			0.5548	1			-0.0896	1		
Rater 3	0.2504	0.4802	1		0.4015	0.351	1		0.139	-0.072	1	
Rater 4	0.175	0.5026	0.4213	1	0.2668	0.2018	0.2873	1	-0.1297	0.1323	-0.2747	1
Synthetic Quality	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**
Rater 1	1				1				1			
Rater 2	0.1058	1			0.5532	1			-0.1991	1		
Rater 3	0.2168	0.4884	1		0.457	0.5291	1		0.3277	0.1729	1	
Rater 4	0.3296	0.4712	0.2651	1	0.4056	0.3049	0.2446	1	-0.0356	0.1146	0.0782	1
Use of Language	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**
Rater 1	1				1				1			
Rater 2	0.1812	1			0.6644	1			0.0254	1		
Rater 3	0.2647	0.5012	1		0.4078	0.2791	1		0.2734	0.1681	1	
Rater 4	0.3157	0.3309	0.477	1	0.3346	0.3145	0.0472	1	0.007	0.2004	0.3468	1
Overall	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**
Rater 1	1				1				1			
Rater 2	0.2178	1			0.5470	1			-0.0068	1		
Rater 3	0.3027	0.5000	1		0.4139	0.3729	1		0.1858	-0.0080	1	
Rater 4	0.3343	0.4661	0.4570	1	0.4066	0.3178	0.2151	1	0.0663	0.2491	0.0500	1

Table 4 Rater Correlation Using AAC&U Rubric

	Book Review				Essay				Research Paper			
Context of and Purpose for Writing	Rater 1	Rater 2	Rater 3	Rater 4	Rater 1	Rater 2	Rater 3	Rater 4	Rater 1	Rater 2	Rater 3	Rater 4
Rater 1	1				1				1			
Rater 2	0.3825	1			0.2663	1			-0.0769	1		
Rater 3	0.2614	0.3823	1		0.2739	0.4291	1		0.0719	0.0958	1	
Rater 4	0.1533	0.0153	-0.0173	1	0.1450	0.2938	0.3871	1	0.1186	0.1581	0.1169	1
Sources and Evidence	Rater 1	Rater 2	Rater 3	Rater 4	Rater 1	Rater 2	Rater 3	Rater 4	Rater 1	Rater 2	Rater 3	Rater 4
Rater 1	1				1				1			
Rater 2	0.4908	1			0.4992	1			0.2505	1		
Rater 3	0.3283	0.1683	1		0.564	0.391	1		0.0340	0.0362	1	
Rater 4	0.3592	0.3824	0.4664	1	0.2035	0.2447	0.2667	1	0.1206	-0.1877	0.0804	1
Genre and Disciplinary Conventions	Rater 1	Rater 2	Rater 3	Rater 4	Rater 1	Rater 2	Rater 3	Rater 4	Rater 1	Rater 2	Rater 3	Rater 4
Rater 1	1				1				1			
Rater 2	0.2657	1			0.4656	1			0.14	1		
Rater 3	0.193	0.4116	1		0.1871	0.2995	1		-0.07	0	1	
Rater 4	-0.0545	0.5185	0.3165	1	0.3138	0.2622	0.3171	1	-0.1267	0.0754	0	1
Content Development	Rater 1	Rater 2	Rater 3	Rater 4	Rater 1	Rater 2	Rater 3	Rater 4	Rater 1	Rater 2	Rater 3	Rater 4
Rater 1	1				1				1			
Rater 2	0.1791	1			0.3613	1			0.2223	1		
Rater 3	0.2722	0.4106	1		0.2829	0.5923	1		-0.1009	0.1430	1	
Rater 4	0.2852	0.3300	0.3649	1	0.384	0.2287	0.2142	1	-0.1771	-0.0928	-0.1176	1
Control of Syntax and Mechanics	Rater 1	Rater 2	Rater 3	Rater 4	Rater 1	Rater 2	Rater 3	Rater 4	Rater 1	Rater 2	Rater 3	Rater 4
Rater 1	1				1				1			
Rater 2	0.4014	1			-0.1435	1			-0.1991	1		
Rater 3	0.1660	0.2338	1		-0.0701	0.4834	1		0.3277	0.1729	1	
Rater 4	0.4082	0.3990	0.1674	1	-0.0829	0.1236	0.2845	1	-0.0356	0.1146	0.0782	1
Overall	Rater 1	Rater 2	Rater 3	Rater 4	Rater 1	Rater 2	Rater 3	Rater 4	Rater 1	Rater 2	Rater 3	Rater 4
Rater 1	1				1				1			
Rater 2	0.3044	1			0.3059	1			0.0802	1		
Rater 3	0.2900	0.3403	1		0.2823	0.4319	1		0.0078	0.0850	1	
Rater 4	0.3048	0.2720	0.2527	1	0.2034	0.2316	0.2851	1	0.0954	-0.0865	-0.0319	1

To ensure validity of the statistic, the level of agreement among raters was measured using Cohen's Kappa statistic using two raters at a time. For each rubric, we compared the six possible combinations or pairings of raters (4 raters examined 2 at a time) to determine the percent agreement within each Performance Area for a particular paper type. The number rater combinations/pairings that had significant agreement out of the 6 possible combinations/pairings and the level of that agreement is displayed in Table 5. Although there is not agreement across all Performance Areas, there is more agreement among the raters on the Academy rubric than on the AAC&U rubric. The only Performance Area that does not have any rater agreement is Context and Purpose for Writing in the AAC&U rubric across all writing types. The Audience-Orientation Performance Area for the Academy rubric also has no rater agreement, except for the Book Review. This finding may be a result of the assignment design. The Book Review specifically included the Audience-Orientation as part of the assignment requirement.

Rater Consistency within the Rubrics

To determine whether the reviewers rated the papers consistently **within** each rubric, in other words, whether all raters scored the papers similarly, a MANOVA, separated by paper type, was run for each Performance Area. For both rubrics there were significant rater effects indicating that **all** raters did not score the papers the same. The raters' scores, however, trended in the same way for each Performance Area. That is, the rater who rated Performance Areas low was always low across all paper types and the rater who rated higher than the others was always higher across paper types. Although all raters did not score the papers the same, this pattern lends itself to the argument that the raters were internally consistent across the paper types on each of the rubrics.

Comparing Across the Rubrics

To facilitate a comparison between the Academy and AAC&U rubrics, pairings were created for each Performance Area based upon the congruence between the subcategories on the Academy rubric and the Capstone descriptions on the AAC&U rubric, as displayed in Table 6.

To determine if these pairings were appropriate and would therefore allow for comparisons across rubrics, we again used Pearson correlation coefficients. The Pearson's correlation coefficient shown in Table 7 indicates the strength of the relationship between the pairings for each paper type[1]. The correlation matrix was created including all Performance Areas. The bolded correlations in Table 7 indicate that the correlation for the Performance Area pairing was stronger than if the Academy rubric Performance Area was paired with any of the other AAC&U performance areas.

Table 6 Performance Areas Pairings

Academy Rubric	AAC&U Rubric
Audience-Orientation	Context of and Purpose for Writing
Discipline Knowledge	Sources and Evidence
Analytical Quality/Critical Thinking	Genre and Disciplinary Conventions
Synthetic Quality	Content Development
Use of Language	Control of Syntax and Mechanics

Rater Consistency across Rubrics

A Generalized Linear Model with repeated measures was used to determine if raters were consistent across

Table 7 Pearson Correlation of Performance Area Pairings

Academy Rubric	AAC&U Rubric	Research Paper	Book Review	Essay
Audience-Orientation	Context of and Purpose for Writing	**0.5494**	0.4406	**0.5425**
Discipline Knowledge	Sources and Evidence	**0.5376**	**0.5815**	**0.5771**
Analytical Quality/Critical Thinking	Genre and Disciplinary Conventions	**0.4513**	**0.5127**	**0.4966**
Synthetic Quality	Content Development	0.4309	**0.5069**	**0.5635**
Use of Language	Control of Syntax and Mechanics	0.3643	**0.3906**	0.4582

NOTE: Bold indicates the Performance Area pairing has highest correlation among all possible pairings.

[1] Spearman rank order correlation coefficients were also calculated and the results were very similar to the Pearson matrices.

rubrics. That is, did the raters score the papers the same on both rubrics? Scores on the five paired Performance Areas were compared across raters for each paper type[1]. The results indicate that the raters did not score the papers the same across rubrics. All of the Performance Area pairings indicate that the raters provided significantly different scores on the research paper. Three of the five Performance Area pairings had significantly different scores for the Book Review, and for the essay, four of the five Performance Area pairings indicated a significant difference in the raters' scores.

Conclusions

Overall, our findings indicate that individually, raters are internally consistent on both rubrics. That is, the rater who tended to rate lower was consistently lower across all Performance Areas. Conversely, the rater who tended to rate higher was consistently higher across Performance Areas. These findings were consistent for all paper types irrespective of which rubric was used. The following table summarizes the statistical measures that we used, their purposes and the empirical results.

The results indicate that the raters did not rate the papers the same way, thus the inter-rater reliability was not strong. However, no training was conducted for raters on the rubrics prior to their utilization. Assumedly, with careful training, the inter-rater reliability would be higher. In fact, the results suggest that the authors who have used the Academy rubric in the past had higher inter-rater reliability. The raters' scores were less dispersed and were shown to have more agreement on the Academy rubric. This pattern was not observed in the use of AAC&U rubric. Thus, it is imperative that raters train on the rubric to yield more consistent results.

Furthermore, the empirical results showed that assignment design and rubric design are two aspects of the same larger process of facilitating learning of how to write in special contexts. For example, Audience Orientation was explicitly discussed in the assignment for the book review, but not in the assignment for the research paper or essay. The scores for this Performance Area were significantly higher for the book review.

Rubric validity is based on instructor judgment in daily use as Cho, et al. (2006) emphasize. Empirical evidence such as positive "concurrent" correlation with other measures with similar constructs and goals, adds a broader foundation for trusting that rubrics can meet scientific standards. Thoughtful collaboration with colleagues in the design and assessment of any rubric is well-advised because many factors, including the wording and meaning of problems, affect the quality of a measure and its uses. The authors recommend that colleagues practice using any rubric with sample papers of varying quality. The research methodology used to examine this rubric provides a model for empirically exploring the reliability and validity of other rubrics as well as for application of the present rubric in new contexts.

Table 8 Summary of Statistical Results

Test	Purpose	Outcome
Cronbach's Alpha	Collapse sub-categories	The individual items are reliable and can be collapsed into a single measure for the Performance Area in the Academy rubric.
Standard Deviation	Dispersion among raters	Across all Performance Areas, the raters' scores were less dispersed when using the Academy rubric than on the AAC&U rubric.
Pearson's Correlation Coefficient	Correlation of ratings across raters	The correlations vary considerably, but generally have a stronger relationship on the Academy rubric than the AAC&U rubric.
Cohen's Kappa	Level of agreement among raters	There is more agreement among the raters on the Academy rubric than on the AAC&U rubric.
MANOVA	Consistency of scores within each rubric	The raters did not score the papers the same, but were internally consistent across the paper types on each of the rubrics.
Pearson's Correlation Coefficient	Correlation between Performance Area pairings	The results vary by paper type and Performance Area pairing, but generally show a correlation between pairings.
Generalized Linear Model	Consistency across rubrics	The raters did not score the papers the same across rubrics

References

Apple, D., Beyerlein, S., Leise, C., & Baeher, M. (2007). Classification of learning skills. In S. Beyerlein, C. Holmes, & D. Apple (Eds.), *Faculty guidebook* (pp. 201-204). Lisle, Il: Pacific Crest.

Association of American Colleges and Universities. (2012). Written communication value rubric. Retrieved April 8, 2012, from http://www.aacu.org/value/rubrics/pdf/WrittenCommunication.pdf

Bruning, R., & Horn, C. (2000). Developing motivation to write. *Educational Psychologist, 35*(1)*,* 25-37.

Burke, K. (2007). Getting student buy-in. In S. Beyerlein, C. Holmes, & D. Apple (Eds.), *Faculty guidebook* (pp. 323-326). Lisle, Il: Pacific Crest.

Burke, K., & Nancarrow, C., (2007). Writing in a disciplinary context. In S. Beyerlein, C. Holmes, & D. Apple (Eds.), *Faculty guidebook*. Lisle, Il: Pacific Crest.

Cho, K., & MacArthur, C. (2011). Learning by reviewing. *Journal of Educational Psychology, 103*(1), 73-84.

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology, 98*(4), 891-901.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-34.

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, *88*(2), 322-328.

Fleiss, J. L. (1971). Measuring the nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378-382.

Graham, S., & Harris, K. R. (2000a). Writing development: Introduction to the special issue. *Educational Psychologist, 35,* 1-2.

Graham, S., & Harris, K. R. (2000b). The role of self-regulation and transcription skills in writing and writing development. *Educational Psychologist, 35,* 3-12.

Hanson, D., & Moog, R. (2007). Process oriented guided-inquiry learning. In S. Beyerlein, C. Holmes, & D. Apple (Eds.), *Faculty guidebook* (pp. 387-390) . Lisle, Il: Pacific Crest.

McCutcheon, D. (2000). Knowledge, processing, and working memory: Implications for a theory of writing. *Educational Psychologist, 35,* 13-24.

Smith, P., & Apple, D. (2007). Overview of quality learning environments. In S. Beyerlein, C. Holmes, & D. Apple (Eds.), *Faculty guidebook* (pp. 311-314). Lisle, Il: Pacific Crest.

Appendix

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 collectios 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

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

Analytic Rubric for Disciplinary Writing Assessment

Date _____ Rater's Last Name _____

Paper # _____ Paper Type _____

Paper Title _____

Performance Area	Assessment *(use the sliders to rate the submission; scale is from 1 to 4)*:				
1. Audience-orientation					
a. Thesis relevance	marginal	adequate	valuable	visionary	1
Check if N/A					
b. Thesis clarity	ambiguous	understandable	well stated	eloquent	1
Check if N/A					
c. Cohesiveness of perspective	absent	very fragmented	somewhat fragmented	artful	1
Check if N/A					
2. Discipline Knowledge					
a. Selection of citations	random	basic	thoughtful	masterful	1
Check if N/A					
b. Depth of disciplinary knowledge	sketchy	fundamental	impressive	profound	1
Check if N/A					
c. Representation of knowledge	rote	sound	substantial	masterful	1
Check if N/A					
3. Analytical Quality/Critical Thinking					
a. Logic of development	unconnected	uneven	well planned	seamless	1
Check if N/A					
b. Validity of evidence	peripheral	limited	acceptable	irrefutable	1
Check if N/A					
c. Application of knowledge	flawed	inconsistent	accurate	innovative	1
Check if N/A					
4. Synthetic quality					
a. Interpretation of evidence	questionable	elementary	sensible	persuasive	1
Check if N/A					
b. Connection to thesis	disjointed	limited	convincing	compelling	1
Check if N/A					
c. Quality of insights/conclusions	simplistic	fundamental	mindful	powerful	1
Check if N/A					
5. Use of Language					
a. Grammar/mechanics	poor	adequate	excellent	flawless	1
Check if N/A					
b. Use of structure (eg: paragraphs/sections)	random	simplistic	appropriate	masterful	1
Check if N/A					
c. Rhetorical eloquence	ineffective	interesting	persuasive	inspirational	1
Check if N/A					

Assess using the SII Method:

Strengths (including why):

Areas for Improvement (including how):

Insights (including significance):

Please share any additional comments or feedback you may have!

Submit Reset

Sample Rating Sequence

	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
A	33	30	22	34	20	26
B	3	22	12	22	32	20
C	17	12	20	20	35	32
B	18	13	4	15	4	21
C	16	13	11	29	15	18
A	29	28	13	25	29	31
C	33	25	24	22	9	2
A	4	12	18	22	1	35
B	10	16	6	26	25	19
B	30	8	27	17	21	9
C	1	3	5	26	34	23
A	32	21	9	17	21	9
C	7	12	16	36	28	19
A	3	36	15	36	35	12
B	16	12	13	36	35	12
A	27	12	17	20	31	7
B	6	31	11	20	31	7
C	30	24	21	31	8	14
C	21	30	10	11	8	23
A	21	23	2	6	13	4
B	27	2	26	24	14	23
A	24	32	17	2	11	19
B	32	1	19	7	29	1
C	32	29	3	27	36	34
B	34	24	5	35	3	10
C	2	9	6	31	19	17
A	35	34	25	33	24	8
A	30	5	10	20	10	26
B	2	17	28	23	28	15
C	26	14	6	27	1	22
B	34	18	36	8	30	14
C	13	25	4	5	15	7
A	23	18	7	31	11	14
C	28	33	10	4	35	18
A	16	28	14	7	5	9
B	33	5	9	29	33	11

A Book Review
B Research Paper
C Essay

Black: Academy Writing Rubric
Red: AAC&U Rubric