

Assessing Students' Writing Performance at the University Level

James Koepfler, Keston Fulcher, and Chris Orem

James Madison University

Assessing Students' Writing Performance at the University Level

Federal and state governments have increased pressure on colleges to demonstrate what students have learned. Relevant to our institution, the State Council of Higher Education for Virginia (SCHEV) has mandated that public colleges and universities report “value-added” data on six general education areas, including writing. This paper focuses on one university’s assessment of writing and how it is collecting validity evidence to support its use for this and other purposes.

While those in and outside of higher education agree about the importance of writing skills in the general education curriculum, the assessment and measurement of this construct has a checkered past (Meredith & Williams, 1984). Often writing is assessed by using student-generated writing samples that are written as part of a time limited, one shot writing test. This approach has been criticized for not capturing the essence of what students learn in the classroom (Wiggins, 1994). Specifically, writing tasks in large scale assessments often fail to give students adequate time to write, do not allow opportunity for revision, and are scored using rubrics that focus heavily on the mechanical components of writing.

Wiggins argued that these assessments are designed to maximize reliability, unfortunately, at the expense of validity. While reliability is no doubt an essential component of validity, without consideration of other components of validity, chasing reliability can actually reduce validity. Therefore, a writing assessment program should endeavor to incorporate a strong program of validity evidence, which attempts to maximize the meaningfulness of scores. Indeed, the Standards for Educational and Psychological Testing (Standards; 1999) acknowledge that, “validity is the most fundamental consideration in developing and evaluating tests. (p.9)”

Piloting New Writing Assessment Procedures at a Midsized University¹

A committee comprised of writing experts across campus explored the university's writing objectives in developing a new writing rubric. The university writing objectives for Cluster One – the general education area encompassing communication – are: (1) Develop and support a relevant and informed thesis, or point of view, that is appropriate for its audience, purpose, and occasion. (2) Analyze and evaluate information to identify its argumentative, credible, and ethical elements. (3) Reflect on civic responsibility as it relates to written discourse (critical thinking, reading, and writing). (4) Demonstrate effective writing skills and processes by employing invention, research, critical analysis and evaluation, and revision for audience, purpose, and occasion. (5) Effectively incorporate and document appropriate sources to support a thesis and effectively utilize the conventions of syntax, grammar, punctuation, and spelling. The University Writing Assessment Committee examined these objectives for guidance to align the rubric with the university writing objectives.

Next, the members of this committee synthesized the most important elements of these objectives, which included 1) usage & mechanics, 2) purpose, 3) organization, 4) style, and 5) complexity. Based on these elements, writing experts created a five trait analytic rubric. For each trait the rubric describes, in behavioral terms, four performance levels: beginning, developing, competent, and advanced. Next, a prompt was designed to elicit the traits identified on the rubric. Students were asked to produce a letter to the student newspaper expressing their response to the question: "Should chronological age (16, 18, 21) be the criteria by which adult responsibilities are granted? Adult responsibilities include issues like: voting, military service, the drinking age, the making of contracts, and legal accountability."

¹ Information presented in this section is also used in a report to SCHEV

The next issue is how to collect information on writing via this rubric. As part of the assessment day system at this institution, students are assessed in a variety of domains, including writing, as incoming students, and then again in their sophomore year when they have earned between 45 and 70 credits. Students are randomly assigned to different assessment rooms. A sample of sophomores was assigned to participate in writing assessment in order to gather validity and reliability evidence for the prompt, writing rubric, and the assessment process.

Methods

Participants

Participants in this study included 40 university sophomores who were randomly selected from a population of approximately 4,000 based on the last 2 digits of their student ID number. The sample of students was 92.5% Caucasian, 2.5% Hispanic, and 5% unspecified, with a mean age of 20.5 ($SD = .64$) years, and a mean grade point average of 3.02 ($SD = .60$). Raters consisted of 28 faculty members from various academic departments. Faculty raters were compensated for training and evaluation.

Measures

University Writing Rubric. Raters evaluated writing samples using the university developed writing analytic rubric (see Appendix A), which focuses on five traits: usage & mechanics, purpose, organization, style, and complexity. For each trait, the rubric describes, in behavioral terms, four performance levels: beginning (1pt), developing (2pts), competent (3pts), and advanced (4pts).

Writing Assessment Evaluation Survey. The survey was developed to assess the conceptual fit between the new writing assessment process and the Universities writing objectives (see Appendix B). The survey assesses four domains of the writing assessment process on a scale ranging from 1 = *Strongly disagree* to 6 = *Strongly agree*. The four domains

assessed include: (1) the ability of the rubric to elicit students writing ability (2) the conceptual fit between the rubric and the writing objectives (3) the usefulness of the rubric in a classroom setting and (4) the effectiveness of the rater training workshop.

Procedures

Students who were assigned to the writing assessment received a hard copy of the writing prompt. The proctor gave students 30 minutes to respond to the prompt in Microsoft Word. Next, students spent 20 minutes taking non-cognitive surveys. Finally, students were instructed by the proctor that they would be given 20 additional minutes to make revisions to the original draft.

These essays were rated by 28 faculty members from various departments. To facilitate rater agreement, faculty underwent a three hour training to apply the rubric similarly to peers. The 40 writing samples were then rated by two raters (a third rater was used to adjudicate discrepancies of two or more points). Last, raters filled out a survey about the writing assessment process, which had items pertaining to the validity of the rubric and the ability of the assessment process to elicit skills implied by the writing objectives.

Results and Discussion

Domain Specification

This pilot study produced some preliminary validity evidence for the writing assessment as well as information that should guide future implementation of this assessment. To begin, the construct of writing competence has been well-defined. This process, according to Benson (1998), is the first step towards building a case for validity. Also known as the “Substantive Stage” (Benson, 1998, p. 12), it is during this phase of collecting validity evidence where experts define aspects of the construct in question, as well as examine the various measures and methods already used to evaluate it. In this study writing competency, as a construct, was defined by

writing experts from multiple disciplines within Cluster One. These experts assembled and established writing objectives along with the analytic rubric (Appendix A).

To gather further evidence about the construct definition, 28 trained raters evaluated the rubric through a survey. The raters, all of whom were faculty members within the Cluster One discipline, provided responses suggesting an alignment of the writing assessment with Cluster One writing objectives. Mean scores of faculty ratings can be found in Appendix C. On average faculty raters *agreed* that the assignment elicited students' skills to develop and support a relevant and informed thesis (Objective 1; $M = 4.93$, $SD = .98$). Faculty raters *slightly agreed* that the assignment elicited students' skills to analyze and evaluate information to identify its argumentative, credible, and ethical elements, as well as, elicited their ability to reflect on civic responsibility as it is related to written discourse (Objectives 2 and 3; $M_s = 3.82$ and 4.11 , $SD_s = 1.8$ and 1.31 ; respectively).

Furthermore, the survey evaluated the level to which raters thought the rubric aligned with the Cluster One writing objectives. Mean scores of faculty ratings can be found in Appendix D. Overall, faculty raters, on average, *agreed* that the rubric corresponded to students' skills to develop and support a relevant and informed thesis (Objective 1; $M = 5.22$, $SD = .64$). Faculty raters', on average, *slightly agreed* that the rubric corresponded to students' skills to analyze and evaluate information to identify its argumentative, credible, and ethical elements, as well as, elicited their ability to reflect on civic responsibility as it is related to written discourse (Objectives 2 and 3; $M_s = 4.11$ and 3.74 , $SD_s = 1.15$ and 1.53 ; respectively).

The third purpose of the survey was to identify if the faculty thought the rubric could be useful in assessing writing within their discipline. On average, faculty raters *agreed* that the rubric could be useful in their discipline ($M = 4.79$, $SD = 1.32$).

The fourth purpose of the survey was to identify the effectiveness of the rater training workshop. Mean scores of faculty ratings can be found in Appendix E. Faculty raters, on average, more than *slightly agreed* that the training session helped them feel more confident rating essays after the training ($M_s = 4.46$ and 4.46 , $SD_s = 1.14$ and $.90$, respectively). Faculty raters, on average, *agreed* that the facilitators were effective in conducting the rater training ($M = 5.15$, $SD = .88$).

Internal Domain Study

Gathering evidence for internal domain studies involves examining empirical evidence to look for response consistencies and to identify the degree to which scores are reliable (Messick, 1995). Benson refers to this step as the “Structural Stage” of gathering validity evidence (Benson, 1998, p. 13). According to Benson (1998), the goal of gathering internal validity is to “determine the extent to which the observed variables covary among themselves, and how they covary with the intended structure of the theoretical domain” (p. 13). One way of measuring the covariance is to examine the amount to which raters agreed with each other on scores for the writing assignment. A high rate of agreement suggests that scores are consistent, and thereby reliable, a core piece of evidence needed for internal validity.

For this study, adjacent and exact rater agreement percentages were calculated. For each essay two raters graded the tests independently and awarded scores according to the rubric. Overall, there was a rate of 40% exact agreement across traits. There was an 86% adjacent plus exact agreement rate among raters, meaning that 86% of the element ratings varied by 1.5 points or less. In the case that scores differed by over 1.5 points, a third rater scored the test.

Discussion

Regarding Benson’s three-step approach to collecting validity evidence, at this point in the process, our strongest evidence is the first step, domain specification. Indeed, local content

experts in writing specified the domain and developed a rubric from this domain. In addition, they approved a prompt to elicit those skills. A separate set of faculty – those that are part of the cluster that will use the rubric – evaluated the rubric and the process by which data are collected and noted a good fit to the cluster-level objectives.

Nonetheless, we are planning to revise the writing assessment evaluation survey to gain more specific information about domain specification. In addition to asking how the rubric and data collection processes relate to division-level goals, we will include items pertaining to the fit of the overall rubric to the construct of writing, each individual element to the construct of writing, and how well the behavioral descriptors within each trait match up to its trait score. In other words, does the description of “Beginning” for the “Organization” trait match what a faculty member believes is appropriate for such a designation.

For Benson’s second step, internal domain studies, our evidence is considerably weaker. For our estimates of reliability, we turned to exact plus adjacent-rater agreement – admittedly, a rough guide to error. That said, the fact that this agreement was over 85% within traits likely indicates that some variability within scores is not error.

To improve our investigation of the internal domain, we have begun data collection on a larger scale. During the week before fall 2009 classes, approximately 200 randomly selected entering first-year students participated in the writing assessment. In the spring of 2010, a sample of sophomores and juniors of about 200 will participate. All 400 essays will be evaluated in May by a team of approximately 30 faculty members who will be calibrated on the rubric. Fifteen teams of two faculty members will each rate approximately 30 papers. Each team’s ratings will be analyzed via generalizability theory with raters and rubric elements treated as facets. The median phi-coefficient of teams will be used as the estimate of reliability for the overall writing scores.

We have yet to evaluate any components of the writing assessment that would be associated with Benson's third step, external domain studies. Of course, this situation is a concern. Toward this end, we plan to investigate group differentiation. These studies include comparing the scores of groups hypothesized to perform differently on the construct. We are planning to compare ratings of entering first-year students who have been exempted from taking the required first-year writing course to those who have not. To receive an exemption, students must achieve a high score on relevant AP or IB exams, possess dual enrollment credit, or make a high SAT Verbal scores plus test out. If exempt students score higher than the non-exempt students then such information would provide initial validity evidence in the external domain.

Even with these planned studies across Benson's three steps, there are two additional major components of this writing assessment that necessitate investigation. The first is our operationalization of revision. Does a twenty minute writing break and subsequent 20 minutes of additional writing really reflect the construct of writing revision? Second, we eventually hope to use this assessment to evaluate the value-added of JMU's general education on students' writing proficiency. Can we be confident that our design is appropriate for this purpose? For example, if we use the same prompt at pre-testing and post-testing, might that cause a practice effect? We hope to tackle these problems in the upcoming year.

In conclusion this pilot study provided preliminary evidence that this writing assessment is appropriate for its purposes. Additionally, it has allowed us to plan more robust studies to collect evidence by which to evaluate the validity of essay ratings. Ultimately, in addition to the accountability question of value added, we hope that data from this burgeoning assessment will provide information to faculty and administrators that will inform JMU's general education curriculum and pedagogy related to writing.

References

- American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association
- Benson, J. J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement, Issues and Practice, 17*(1), 10-17.
- Meredith, V. H. & Williams, P. L. (1984). Issues in direct writing assessment: Problem identification and control. *Educational Measurement, Issues and Practice, 3*(1), 11-15.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.
- Wiggins, G. (1994). The constant danger of sacrificing validity to reliability: Making writing assessment serve writers. *Assessing Writing, 1*(1), 129.
- .

Appendix A
Writing Rubric

JAMES MADISON UNIVERSITY WRITING RUBRIC					
<i>2008-2009</i>					
TRAITS	Beginning (1pt)	Developing (2pts)	Competent (3pts)	Advanced (4pts)	Score
<p><u>Usage & Mechanics:</u> <i>Generally includes issues dealing with writing conventions. Features considered may include clarity, sentence structure, grammar, spelling, punctuation, and capitalization.</i></p>	<p>Contains pervasive errors in mechanics, usage, grammar, or sentence structure. Problems interfere with meaning or distract the reader.</p>	<p>Contains some errors in mechanics, usage, grammar, or sentence structure. Problems may, on occasion, compromise meaning or distract the reader.</p>	<p>Is generally free of errors in mechanics, usage, grammar, or sentence structure. Reads smoothly. Problems do not compromise meaning.</p>	<p>Demonstrates mastery of spelling, punctuation, usage, and mechanics. May use language and punctuation to enhance meaning.</p>	
<p><u>Purpose:</u> <i>Generally refers to conveying a message appropriate to its audience. Features may include a thesis or central idea, topic selection, relevance, clarity, and focus.</i></p>	<p>Inappropriate for the audience, or intended audience unclear. Lacks a central idea, thesis, or goal, or these elements are unfocused, random, or confusing.</p>	<p>Occasionally appropriate for the audience or intended audience somewhat clear. Central idea, thesis, or goal emerges but may lack focus or consistency.</p>	<p>Mostly appropriate for a defined audience. Exhibits a generally clear and consistent central idea, thesis, or goal.</p>	<p>Clearly appropriate for a well-defined audience. Consistently exhibits a focused central idea, thesis or goal.</p>	
<p><u>Organization:</u> <i>Generally refers to the coherence of the writing. Features may include appropriate format, balance and ordering of ideas, flow, and transitions.</i></p>	<p>Lacks a sense of overall structure; no sense of beginning, middle, or end. No paragraphs or division into paragraphs lacks logic. Lacks transitional words, phrases, and sentences between or within paragraphs.</p>	<p>Contains an overall sense of beginning, middle and end, but paragraph sequence may be confusing. The order or balance of ideas within paragraphs is inconsistent. Little or inappropriate use of transitions.</p>	<p>Effective structure and arrangement of ideas. Order of paragraphs may, occasionally, appear mechanical or awkward. Order or balance of ideas within paragraphs is generally consistent and cohesive. Transitions present but may be cumbersome or repetitive.</p>	<p>Rational, sensible, and deliberate structure that enhances and clarifies meaning. Transitions show relationships among ideas.</p>	

JAMES MADISON UNIVERSITY WRITING RUBRIC

2008-2009

TRAITS	Beginning (1pt)	Developing (2pts)	Competent (3pts)	Advanced (4pts)	Score
<p><u>Style:</u> <i>Generally refers to the choices the writer makes for specific audiences. This may include features like tone, sentence length and structure, phrasing, and word choice.</i></p>	<p>Writing has an inappropriate tone.</p> <p>The sentences and phrases are simplistic, unvaried, or wordy. Writing is stiff, awkward, and difficult to follow.</p> <p>Unclear or incorrect use of terminology or vocabulary.</p>	<p>Writing has an inconsistent or occasionally inappropriate tone.</p> <p>Some sentences and phrases are repetitive, bland, or awkward. Writing is occasionally difficult to follow.</p> <p>Some misused terminology or vocabulary. Word choice may be ineffective.</p>	<p>Writing has a consistent and appropriate tone.</p> <p>Sentences and phrases are typically concise and effective but may be somewhat mechanical. Writing is easy to follow.</p> <p>Terminology or vocabulary is appropriate and sensible but may be predictable.</p>	<p>Tone contributes to reader comprehension.</p> <p>Uses varied sentence structure and phrases to convey meaning and to create interest and engagement.</p> <p>Vocabulary is sophisticated, precise, and varied.</p>	
<p><u>Complexity:</u> <i>Generally refers to depth or sophistication of thoughts and ideas. Features may include research, reasoning, evidence, detail, development, creativity, originality, integration, and perspective.</i></p>	<p>Reasoning is uncritical, illogical, superficial, or simplistic.</p> <p>No evidence or inaccurate and/or inappropriate evidence. Fails to cite or utilize sources. Fails to consider alternative viewpoints.</p> <p>Perspective is one-dimensional, offering only generalizations and stereotypical points.</p>	<p>Reasoning may be faulty or inconsistent.</p> <p>Evidence may be overly general, misinterpreted or misapplied. Insufficient use of sources. Limited consideration of alternative viewpoints.</p> <p>Tends to borrow or simply summarize the perspectives or arguments of others without integration.</p>	<p>Reasoning is logical and consistent.</p> <p>Evidence is appropriate, and, for the most part, effective. Moderate support from acceptable sources. Some consideration of alternative viewpoints.</p> <p>Clearly understands and integrates perspective or arguments of others.</p>	<p>Reasoning demonstrates depth and sophistication of thought.</p> <p>Point of view or argument well-reasoned, balanced, and supported with specific details, facts, and evidence synthesized from well-chosen sources.</p> <p>Perspective or analysis is fresh, original, or insightful.</p>	

Appendix B
Writing Assessment Evaluation Survey

Dear rater training participants,

As part of this session we would appreciate your input via this short survey. The survey is divided into four parts: (1) How well does the assessment day writing assignment relate to each of the Cluster 1 Writing Objectives? (2) How closely is the rubric aligned with the Cluster 1 Writing Objectives? (3) How useful do you think this rubric will be in your area? And, (4) General evaluation of the training session

- I. The assessment day writing assignment consisted of a prompt regarding one’s views on the wisdom of using age as the criterion by which Americans are granted access to public rights (e.g., driving, drafting into the military, drinking). Students were given 30 minutes to write an essay. Then, after 20 minutes of non-cognitive assessment, they were allowed 25 minutes to revise.

Please indicate your level of agreement with the following statements by putting check marks in the appropriate boxes.

The assessment day writing assignment elicited students’ skills to:

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
Develop and support a relevant and informed thesis, or point of view, that is appropriate for its audience, purpose, and occasion						
Analyze and evaluate information to identify its argumentative, credible, and ethical elements.						
Reflect on civic responsibility as it relates to written discourse (critical thinking, reading, and writing).						

- II. The rubric is what you just used to evaluate several papers.

Please indicate your level of agreement with the following statements by putting check marks in the appropriate boxes.

What the rubric evaluates corresponds to the following objectives:

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
Develop and support a relevant and informed thesis, or point of view, that is appropriate for its audience, purpose, and occasion						
Analyze and evaluate information to identify its argumentative, credible, and ethical elements.						
Reflect on civic responsibility as it relates to written discourse (critical thinking, reading, and writing).						

III. Usefulness to the papers you grade in your discipline.

Please indicate your level of agreement with the following statement by putting check marks in the appropriate box.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
This rubric could be used effectively to evaluate papers in my discipline.						

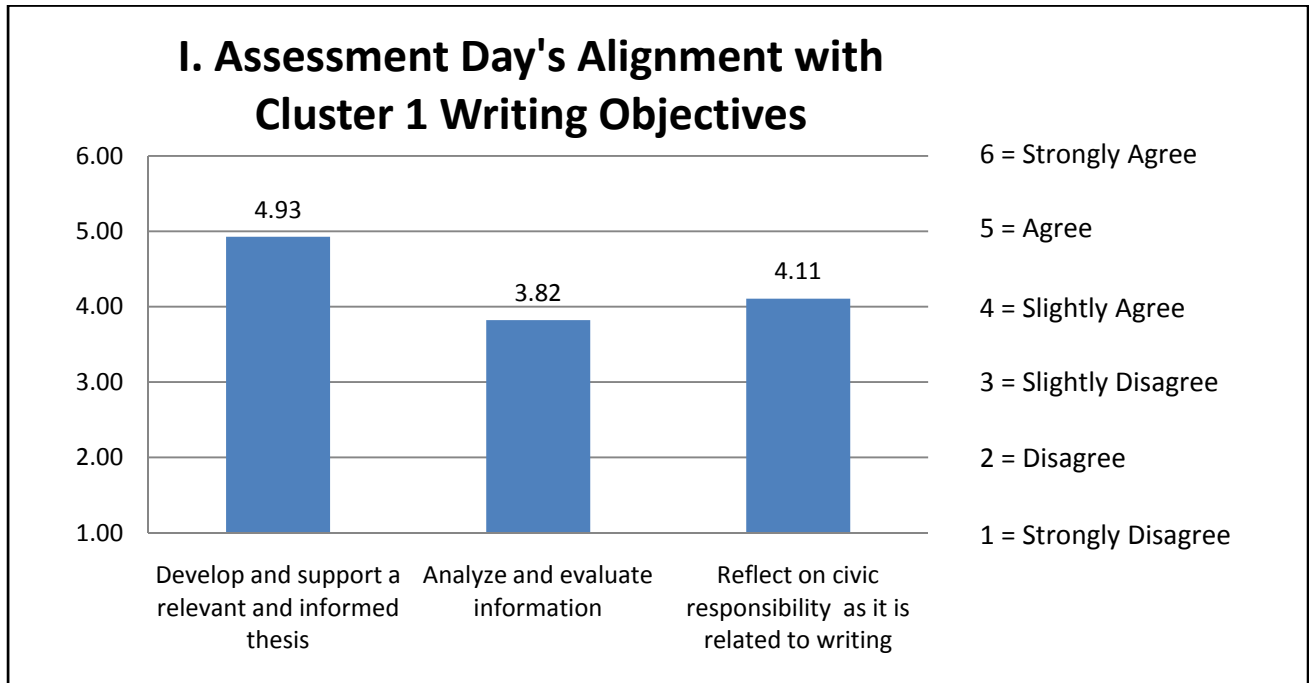
IV. Effectiveness of the workshop

Please indicate your level of agreement with the following statements by putting check marks in the appropriate boxes.

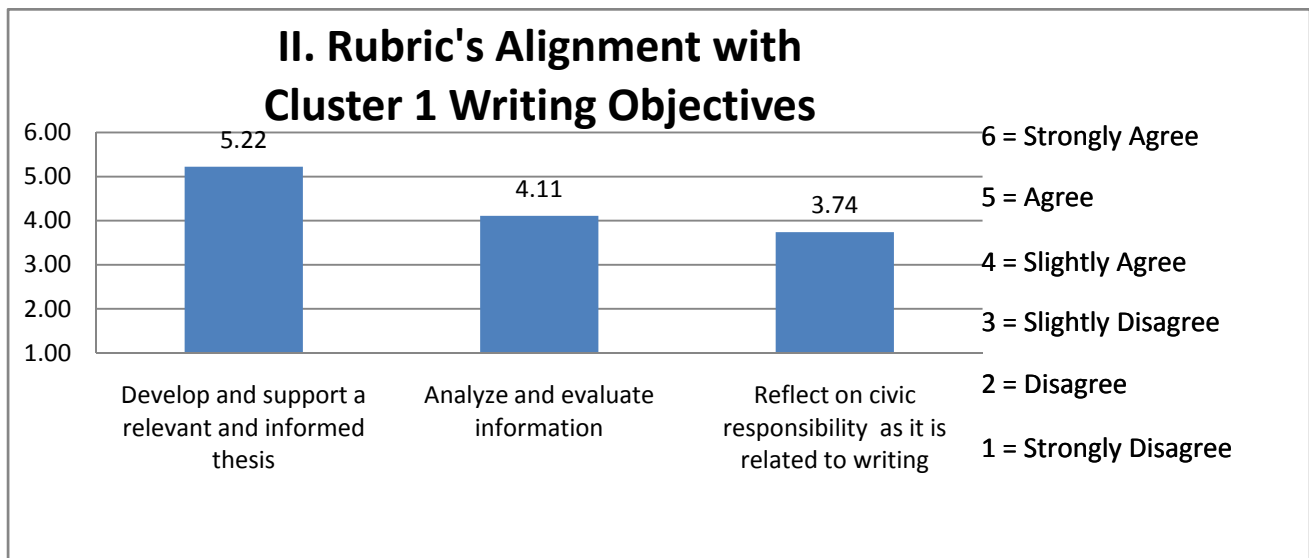
	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
I feel confident evaluating papers using this rubric.						
I am consistent with my peers when rating essays using this rubric.						
The facilitators were effective in the rater training.						

Please provide any other comments regarding the student assignment, the rubric, or the rubric's potential use in your classes, and the rater training.

Appendix C
Mean Scores of Faculty Ratings



Appendix D
Mean Scores of Faculty Ratings



Appendix E
Mean Scores of Faculty Ratings

