2015-12-01

Evidences of Critical Thinking in the Writing of First-Year College Students

Shannon Bryn Soper
Brigham Young University

Evidences of Critical Thinking in the Writing of First-Year College Students

Shannon Bryn Soper

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Arts

Kristine Hansen, Chair
Brian David Jackson
Deborah Marquiss Dean

Department of English

Brigham Young University

December 2015

ABSTRACT

Evidences of Critical Thinking in the Writing of First-Year College Students

Shannon Bryn Soper
Department of English, BYU
Master of Arts

A healthy civil society depends on citizens who have mature critical thinking skills and a willingness to entertain opposing points of view. The development of critical thinking in young adults has long been studied, but there has been little agreement on what the attributes of critical thinking are and how to reliably assess them. While many studies have attempted to assess the critical thinking abilities of college students, none have yet measured critical thinking through using the Critical Thinking Analytic Rubric (CTAR) to assess first-year college students' writing. This study used a modified version of the CTAR rubric to investigate students' critical thinking in writing completed for an American Heritage course. Four hypotheses were tested: (1) that raters would use the rubric with high inter-rater reliability estimates; (2) that there would be a significant relationship between the scores from the earlier holistic rubric used in the 2015 Hansen et al. study and the scores from the analytic rubric used in this study; (3) that there would be a significant relationship between analytic scores and ACT and GPA scores; (4) that there would be a significant relationship between essay score and gender. Findings included the following: (1) The inter-rater reliability for the overall scores of the papers was 0.898, which exceeds the 0.70 acceptable level. However, the inter-rater reliability for sub-scores was negative and required further investigation. (2) There was no significant relationship between the scores of the Hansen et al. study and this study. (3) There was no significant relationship between essay scores and ACT and GPA scores. (4) There was a significant relationship between essay scores and gender, with female students scoring higher than male students.

Keywords: composition, assessment, critical thinking, first-year writing, scoring rubrics, analytic rubric, BYU, American Heritage

ACKNOWLEDGMENTS

Many wonderful people have supported the completion of this project, including but not limited to the following. My thesis chair, Dr. Kristine Hansen, has been the best mentor I could ask for. She coached me as I learned to think and write in a genre that was new to me: quantitative composition studies. She also encouraged me when I encountered what I perceived to be dead-ends and went above and beyond her responsibilities in helping me become a better researcher. The project had some unexpected twists and turns, and she was a constant source of support and motivation throughout the experience. Dr. Brian Jackson and Dr. Debbie Dean also responded with warmth, enthusiasm, and insight as they provided prompt and thorough feedback during different stages of the project. Dr. Dennis Eggett also took a special interest in helping this project come to fruition by not only helping me gather and process statistically sound data, but also gather more data when the first set showed some inconsistencies. Finally, my family and friends were incredibly supportive of me throughout this process and always supported my desire to press forward with my education. Last but certainly not least, I want to thank God for the creation and inspiration of all good things that happened in this project; He is the source of my desire and motivation to press forward.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Moral psychologist Jacob Haidt, in his book *The Righteous Mind*, points to a human behavior that may explain the lack of civility in political dialogue today. Haidt observes that we human beings "make our first judgments rapidly, and we are dreadful at seeking out evidence that might disconfirm those initial judgments," or, in other words, we have what's called a "confirmation bias" (55). For whatever reason, we avoid thinking critically about our own beliefs, and merely ignore or reject the beliefs of those who disagree with us. This tendency towards a "confirmation bias" may contribute to the often rancorous and uncompromising political atmosphere today, and it likely contributes to gridlocked discourse in many arenas.

A review of literature in psychology, critical thinking, and writing studies suggests how people might overcome their stubborn death-grip on first judgments through critical thinking; however, this same review of literature reveals gaps in what we know about college students' critical thinking, even though many universities aim to empower students to participate in civil dialogue. For example, psychologist Kohlberg found that children grow into mature moral thinkers through "frequent opportunities for role taking—for putting themselves into another person's shoes and looking at a problem from that person's perspective" (9). His findings are evocative, yet cannot be applied to college-aged students without further study. Other psychologists outline myriad metrics for measuring critical thinking, many of which have not been tested rigorously at the college level across disciplines or subject matters. Many leading educational psychologists, such as Bloom, Piaget, Perry, Kohlberg and Gilligan, developed models for measuring critical thinking in the 1960's and 70's, but few studies have attempted to replicate or confirm their findings empirically. Some of the subskills of critical thinking, among others, are the ability to perceive "irony, ambiguity, and multiplicity of meanings or points of

view" as well as "the development of open-mindedness, reciprocity. . . and autonomous thought" (Lazere 2). While these psychologists developed foundational models for conceptualizing and defining critical thinking, few large analytical studies have used those criteria to attempt to measure the critical thinking of students today.

Second, many critical thinking studies have produced clear results, but have not gathered data about writing studies, and the data and resulting advice on what to do to improve critical thinking in one field may not apply in other fields. According to a study by the Center for Teaching Excellence at the University of Arkansas at Little Rock, teachers need a way to measure critical thinking that is tailored to their discipline, because theoretical or abstract measurements are not sufficient (Thaxton). Lazere agrees with Thaxton; he gathered a significant body of references to studies that indicate that "neither critical thinking nor cognitive development can effectively advance except in dialectical interaction with a substantial body of domain-specific knowledge" (3). Examples of studies that have produced clear results but fail both Thaxton and Lazere's criteria of being specific to the writing domain include the studies of McGuire and Pinos Beck. McGuire found that through direct rhetorical analysis—using argument mapping, Thinker's Guides (based on Paul's model of critical thinking), and Socratic questioning—college students improved their critical thinking ratings on the California Critical Thinking Skills Test 2000. However, this test does not use writing but rather multiple-choice test questions. Furthermore, Pinos Beck's findings on first-year nursing students' critical thinking skills is thorough and reports both quantitative and qualitative data, but the findings on nursing students are not yet proven to be applicable to other disciplines.

Third, and finally, the critical thinking studies in the writing studies discipline reveal gaps; some studies rely on anecdotal evidence rather than data, while others focus on high school

students but not college students. Karras advises high school teachers that high school students will develop better critical thinking skills if the teacher shows them how to write highly structured, even rigid problem-centered essays in high school history courses. Karras has not yet gathered data on the effectiveness of this technique. Shafer also counsels community college professors to challenge their students to write in response to the more difficult questions of democracy while avoiding the five-paragraph essay format, although teachers have countered by saying that the five-paragraph structure helps students perform better than they would without the structure. Shafer also did not gather data on his technique. Furthermore, Karras, Shafer, and Pennell all point to the basic five-paragraph high school essay as the sort of model that, ironically, may actually prevent first-year college students from thinking more critically in their writing, but they fail to gather data to prove their assertions.

Other scholars, such as Nussbaum, have suggestions for facilitating critical thinking, but, again, have only gathered anecdotal evidence of their solution's effectiveness. Nussbaum recommends teaching students to structure their arguments using vee-diagrams, which are essentially a way to weigh the strength of arguments and counterarguments, but he did not measure the effectiveness of this technique. Pennell proposed an assignment in which he asked students to respond critically to each source in their bibliographies for a certain paper, but he did not publish the measured effectiveness of his assignment. In a high school Shakespeare unit, Strom found that performance-based learning increased critical thinking skills more than seat-based learning (by which he means listening to lectures, reading, writing, and test-taking). While Strom's study gathered sound data, no one has yet proved that his findings are transferrable to college-aged students.

Because of the aforementioned gaps in existing literature, this study measures the critical thinking of college age students through their writing, by using an analytic rubric. I chose to study first-year students at Brigham Young University (BYU) not only because there was a gap in existing literature on college-aged students' critical thinking, but also because the data is likely to be valuable to many universities and professional organizations (CWPA, NCTE, NWP) that have critical thinking as an outcome. BYU, like many universities, wants its students to become the kind of balanced individuals who will strengthen their communities through civic duty (BYU Mission Statement). As part of this aim, the university instituted the American Heritage General Education requirement, a class that aims to help students understand the political, economic, and philosophical principles that undergird American culture and government. The aim is to have students leave the class "better informed and prepared to make a meaningful contribution to the world" (American Heritage Course Learning Outcomes). In some sections of the course, instructors assign writing tasks to help students reach the class objectives.

For the present study, and based on available funding, I hired three raters to read 30 randomly selected essays written by AH students in fall 2012, which already had been holistically evaluated in a previous study of first-year students' writing ability (see Hansen et al.). These student essays were a valuable source for assessing students' critical thinking skills because, as Bean argues, "thesis-governed writing entails a complex view of knowledge in which differing views about the nature of truth compete for allegiance" (Bean 22). Bean's observation is particularly applicable to the kind of writing the American Heritage students were asked to do (see Appendix A for the prompt). The prompt specifically asked the students to take a stance on a central question of democracy, which required them to argue and rebut counterarguments—all essential skills in critical thinking.

Finally, I modeled my rubric after Saxton et al.'s Critical Thinking Analytic Rubric (CTAR), which they used to analyze the writing of 304 high school students in the initial study (Saxton 257; see Appendix B). The CTAR is the first of its kind, and seemed to be a good instrument for assessing students' critical thinking abilities for three reasons. First, it is based on the Delphi definition of critical thinking, a definition that was formed at a meeting sponsored by the American Philosophical Association in the late 1980's (Saxton 254-5) and which the APA still relies on as its definition for critical thinking. Second, Saxton et al. had already used the Critical Thinking Analytic Rubric (CTAR) to analyze the writing of 304 high school students (who also participated in concurrent college enrollment) with acceptable levels of inter-rater reliability (≥ 0.70) in all rubric categories (Saxton 257, 251; see Appendix B).

Third, an analytic rubric is necessary in this study to produce the kind of analytic data that the psychology, critical thinking, and writing studies disciplines are missing: exact measurements on the critical thinking abilities of college-level writers. According to Saxton et al., "the need for instruments that measure critical thinking sub-skills, rather than critical thinking in general, argues in favor of the use of analytic rubrics rather than holistic rubrics" (253). Because of the previous three reasons, the CTAR rubric was a model for me as I created my own instrument for measuring critical thinking. I saw a need to test an analytic rubric, similar to the CTAR rubric, with first-year college student writing.

In the current study, four hypotheses were tested to fill the gaps in literature in the psychology, critical thinking, and writing studies disciplines. In summary, those disciplines are missing concrete studies that have gathered analytic data on the critical thinking of college student writing. First, I hypothesized that raters in this study would be able to use an analytic rubric to produce consistent ratings of student critical thinking in the essays. Second, I

hypothesized that the papers that scored high on the earlier holistic rubric used in the Hansen et al. study would also score high on the analytic critical thinking rubric because I guessed that either the holistic rubric used by Hansen et al. was actually measuring critical thinking or that critical thinking would correlate with a holistic score. Third, I hypothesized there would be a correlation between student GPA and ACT scores and the overall essay score given by raters in this study. Fourth, I hypothesized a correlation between gender and score, because prior studies suggested this correlation as a possibility (Willingham et al., and Lane).

METHODS

Definitions

In order to conduct this research, the first task was to define and find a way to measure critical thinking—not a simple endeavor. Bean asserts that experts agree on what critical thinkers should be able to do: "demand justification of claims, seek to disconfirm hypotheses, avoid hasty conclusions, and provide reasons and evidence for their own claims" (Bean 20). However, Bean seems to be ignoring other critical thinking definitions that differ from his; for example, the Delphi definition claims that the six essential cognitive skills in critical thinking are the ability to interpret, analyze, evaluate, integrate, infer, and explain (Saxton 254-5). We see variations in definitions of critical thinking, but theorists and previous studies disagree much more drastically on how to operationalize critical thinking abilities. In other words, not only do experts disagree on the definition of critical thinking. They also disagree, to a greater degree, about how to create an instrument that allows evaluators to objectively assess the attributes of critical thinking.

Researchers assess critical thinking in student writing differently now than they have in the past; in the 60's, some people thought one could measure the general level of complexity in writing, and thereby know the level of critical thinking. Many measured syntactic maturity by

sentence length, clause length, subordination ratio, and kinds of subordinate clauses (Hunt 1). However, Hunt determined that measuring the length of the T-unit, the combination of a main clause with all associated subordinate clauses, was a more accurate way to determine syntactic maturity (20-21).

In the 70's and 80's, Moffett, among others, expanded their definition of critical thinking to include not only the sentence complexity, but also the writer's stance towards their audience. Moffett identified linguistic codes that mark developmental shifts, including the shift from addressing oneself as an audience to addressing an unknown audience, which marks the developmental shift from egocentrism to being other-focused, a shift that Piaget insisted was necessary for critical thinking (24, 57). In the 90's and continuing into the present, many definitions of critical thinking proliferate, but two important and competing definitions are important to note. Camp asserts that some researchers insist that critical thinking is context-specific while others claim that critical thinking is a general skill that applies across contexts. Yet other researchers, including Beaufort, support the definition that critical thinking consists of both context-specific *and* generalizable skills. Beaufort's five domains are an example of a guide that outlines five general categories within which a writer must develop context-specific knowledge in order to become an expert, which Camp asserts is the same as becoming a critical thinker (Camp 100, Beaufort 19).

In the current study, I modeled my definition of "mature critical thinking" after the Saxton et al. study's definition, which is the Delphi definition. I define "mature critical thinking" as thinking that rates high on all of the attributes in the CTAR rubric: interpretation, analysis, evaluation, inference, explanation, and disposition.

Instrument

The instrument I developed for scoring the papers was an analytic rubric that includes all of the six sub-skills of critical thinking that Saxton et al used in their rubric, although I combined two of the sub-skills into one sub-skill (Saxton 245, Appendix C). I initially planned on using the full CTAR rubric, but upon initial testing, realized that the sub-score descriptions were too long and detailed for the raters to read and apply in this study. It would have been too time consuming for my purposes and for the time and funding I had available to use. I used the same categories but modified the CTAR rubric as follows (see CTAR Rubric in Appendix B and Soper Rubric in Appendix C):

1. At first I combined the sub-categories Interpretation and Analysis into one category because the definitions seemed so similar. However, during initial scoring exercises, the raters expressed that they wanted to separate Interpretation" and Analysis into different categories and instead combine the Inference and Explanation categories into one category called "Evidence." I complied with their request this because during initial scoring exercises the skills seemed so interrelated that the differences between them were too negligible to warrant having separate categories.
2. I reduced the scale from 0-6 points to 1-4 points in order to simplify the grading process for the raters. A six-point scale seemed too broad and called for judgments that were too fine for raters to make. A four-point scale seemed clearer because the following phrases could be used to represent each point:
    a. Fails to do this = 1 point
    b. Begins to do this = 2 points
    c. Does this adequately = 3 points

d. Exceeds expectations = 4 points

3. I shortened the explanation of each subcategory to one sentence. The CTAR rubric has longer explanations.

Table 1 below compares the original CTAR rubric with my modification of it. Please see Appendix B for the CTAR rubric and Appendix C for the full Soper rubric that includes definitions of all critical thinking attributes.

Table 1. Comparison of CTAR rubric and Soper rubric

CTAR Rubric		Soper Rubric	
Subcategories	Points Possible	Subcategories	Points Possible
Interpretation	0-6	Interpretation	1-4
Analysis	0-6	Analysis	1-4
Evaluation	0-6	Evaluation	1-4
Inference	0-6	Evidence (combined Inference & Explanation)	1-4
Explanation	0-6		
Disposition	0-6	Disposition	1-4

Objects of Study

From the previous study (see Hansen et al.), I chose 15 high-scoring essays and 15 low-scoring essays, from which all identifying information was deleted. I chose to study American Heritage papers for three reasons: First, the professors and students in the 2012 fall semester course had already agreed to participate in a large-scale composition study, so the papers had already been gathered, and I received permission from BYU's IRB to use a subset of the papers as secondary data. Second, the prompt the students wrote to required them to address counter-arguments, which gave me a way of assessing whether the students were able to think critically about others' positions in the way that mature citizens should. Third, many researchers agree that only through studying a specific subject matter long enough can a student amass a depth of knowledge that is deep enough to enable rich critical thinking (Lazere). The writing produced in the American Heritage course at BYU is one of the best sources of data for measuring the critical

thinking of freshmen because students spend an entire semester pondering and learning about the purposes and philosophies of government in the American democracy.

Quantitative Procedures

Three MA students in the English department were recommended as excellent raters by the head of the University Writing Program, Dr. Brian Jackson. They were provided compensation for rating the papers through the Thayer Research Award from BYU.

The raters used my shortened version (see Appendix C) of the CTAR rubric (see Appendix B) to rate the papers. They attended a two-hour training during which they graded sample papers and discussed the rubric. The inter-rater agreement during this grade-norming exercise was good, and the raters appeared to need no further instruction. The only feedback they gave was that the sub-category "Interpretation & Analysis," which I had at first combined for the purpose of simplifying the rubric, should be separated into two categories. Both Rater 1 and Rater 2 commented that the combination was confusing enough that they couldn't determine the score, as you'll see in table 2 below where they marked "NA." I participated in grade-norming as Rater 4.

Table 2. Grade-norming exercise scores

	Paper #62				Paper #132				Paper #218			
Category	Rater 1	Rater 2	Rater 3	Rater 4	Rater 1	Rater 2	Rater 3	Rater 4	Rater 1	Rater 2	Rater 3	Rater 4
Interpretation & Analysis	1	3	3	2	3	[NA]	3	3	[NA]	1	2	2
Evaluation	1	1	2	1	2	1	3	3	2	2	2	2
Evidence	1	2	2	1	3	2	3	2	3	2	2	2
Disposition	1	1	2	1	3	3	2	3	1	2	2	1
Overall Score	4	7	9	5	11	6	11	11	6	7	8	7

Two days later, the raters gathered, and, after a brief review of the purpose of the study and the schedule for the day, they began grading. I purchased a pack of colored highlighters for each rater with no specific intentions for how they would use the highlighters. Shortly after they

began the grading, the raters came up with the idea of using a different color to mark passages in each essay that they thought corresponded to each sub-category in the rubric. For example, they agreed to use blue to mark passages that displayed evidence of the disposition of the writer. This unplanned procedure turned out to be a very fortuitous step in the rating because it later allowed me to see patterns of colors in each paper, which may relate to patterns of critical thinking. Because each rater had their own photocopied set of papers, their written comments and highlighting did not influence other raters. The raters took approximately 20 minutes to rate each paper, spending a total of 5 hours, with a short lunch break in the middle. At the end of the day, I entered the ratings in a spreadsheet. When there was a disagreement of four points or more on a paper's overall score, we had a third rater score it, based on the recommendation of our statistics consultant, Dr. Dennis Eggett. Finally, we discussed their initial impressions, and later, we interviewed them to find out more about the rater's rating process.

Qualitative Procedures

In Table 4 below, it becomes obvious although the inter-rater reliability was high for overall paper scores, there were startling negative inter-rater reliability scores in the raters' sub-scores on the analytic rubric. It was determined a second qualitative step needed to be added to the methods in order to attempt to explain the very unusual findings. After getting the quantitative results, I amended the IRB protocol to include an interview with the raters to try to determine what they were thinking as they made judgments about the sub-scores they gave to each paper. When each rater arrived at the interview, I presented each rater with three or four papers that he or she rated during the quantitative portion of the study. I requested that they review each paper, its accompanying rubric, and all their original markings and notes on the documents. Susan (this and other names are pseudonyms) reviewed papers #43, #56, #11. Derek

reviewed #144, #57, #83, #46. Jacob reviewed papers #43, #144, #56, #46. I selected papers for which the raters' overall scores differed by four or more points *and* papers upon which the raters' overall scores were within a point or two of each other. I also chose papers that represented high, medium, and low ratings for each rater. I chose these papers in order to find out any possible reasons for both disagreement and agreement in rating.

The raters answered four questions:

1. After reviewing these three or four papers that you already graded, what do you remember about your thought process when grading each one? Why did you decide to score it the way you did?
2. What patterns did you notice in papers that scored low? High?
3. What factors influenced you that aren't specifically mentioned in the rubric?
4. How did your understanding of the rubric change as time went on?

RESULTS

Quantitative Findings

Table 3 summarizes the main findings. See Appendix D for the full table of results, including sub-scores and correlations with Hansen et al. scores, ACT and GPA, and gender. Table 3 shows that the lowest scoring paper received 5 out of 20, while the highest scoring paper received 20 out of 20. Anytime that rater 1 and rater 2 disagreed by more than four points, a third rater scored the paper (see Appendix D).

Table 3. Rating results

ID	Rater 1	Overall Score 1	Rater 2	Overall Score 2
Paper ID 83	Jacob	14	Devon	7
Paper ID 56	Shelli	7	Jonathan	5
Paper ID 15	Shelli	9	Jacob	11
Paper ID 19	Shelli	9	Devon	15
Paper ID 89	Shelli	8	Devon	10
Paper ID 99	Jacob	17	Derek	12
Paper ID 51	Jacob	12	Derek	12
Paper ID 76	Susan	10	Derek	13
Paper ID 107	Susan	10	Jacob	10
Paper ID 7	Susan	11	Jacob	12
Paper ID 57	Jacob	20	Derek	20
Paper ID 69	Susan	19	Jacob	15
Paper ID 135	Jacob	13	Derek	10
Paper ID 143	Susan	8	Derek	14
Paper ID 2	Susan	16	Derek	10
Paper ID 11	Susan	20	Derek	9
Paper ID 43	Susan	13	Jacob	20
Paper ID 58	Susan	14	Derek	20
Paper ID 59	Susan	11	Jacob	18
Paper ID 46	Jacob	20	Derek	12
Paper ID 88	Susan	12	Jacob	9
Paper ID 122	Susan	10	Derek	16
Paper ID 144	Jacob	10	Derek	16
Paper ID 35	Jacob	20	Derek	17
Paper ID 37	Susan	8	Derek	12
Paper ID 67	Susan	13	Jacob	13
Paper ID 5	Jacob	13	Derek	11
Paper ID 52	Susan	17	Derek	15
Paper ID 65	Jacob	9	Derek	8
Paper ID 142	Susan	8	Jacob	11

The first finding was that the raters were very much alike in their overall scoring of papers. Prof. Dennis Eggett and I calculated the inter-rater reliability estimates by subtracting the variance between raters from the variance between students, and dividing that number by the

variance between subjects.

((variance between students)–(variance between raters for the students))/

÷ variance between students

= inter-rater reliability estimate

The inter-rater reliability for their overall scores was 0.898, which far exceeds the 0.70 acceptable level (see table 4). This means that the raters essentially agreed about what overall score each paper should get. However, the agreement between raters on the sub-scores was very poor; in fact, the inter-rater reliability scores were negative. This unusual finding is what we might expect for a randomized rubric (i.e., if we scrambled the order of the five sub-scores in each rater's rubric).

Table 4. Inter-rater reliability of scores between judges

Rubric categories	Variance Between Papers (rater_Papers)	Variance Between Raters for the students (Residual)	Inter-rater Reliability Estimates (N = 30)
Overall Score	8.8949	0.9074	0.89798
Interpretation	0.1395	0.3102	-1.22414
Analysis	0.2692	0.4014	-0.49133
Evaluation	0.3227	0.4292	-0.32990
Evidence	0.1710	0.3368	-0.96947
Disposition	0.2320	0.3794	-0.63550

The second finding was that there was no significant relationship between the Hansen et al. scores with the current study's average scores (0.22 correlation, $p = 0.25$). The third finding was that there was no significant relationship between students' GPA and their paper's score (0.016 correlation, $p = 0.93$) and no significant correlation between students' ACT scores and their paper's score (0.188 correlation with ACT, $p = 0.32$). The fourth finding was that gender was significantly related to the overall scores on the paper ($p = 0.03$), as women students scored higher on the overall score than men, on average. Females, on average, got an overall score of 14/20 while males scored 10/20 on average (see fig. 1).

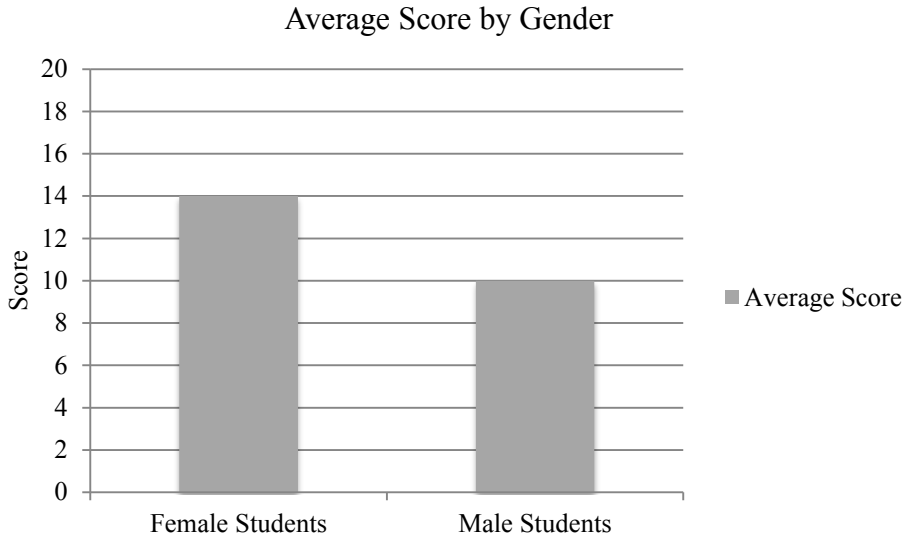

Figure 1. Average score by gender

An additional finding, one that is difficult to express quantitatively, was that those students who scored highest overall exhibited a different pattern in their critical thinking traits than those students who scored lower. According to the color-coding the raters did as they read the essays, the high scorers wove counterarguments, analysis, and explanations throughout each paragraph. The low scorers, on the other hand, tended to address each facet of critical thinking in chunks; one paragraph devoted to their argument, the next paragraph to the counterargument, the next to an analysis, and so on. This generalization is based on a visual inspection of the color-coding the raters did rather than on specific measurements.

Qualitative Findings from Interviews with Raters

The first question the raters answered is: "After reviewing these three or four papers that you already graded, what do you remember about your thought process when grading each one? Why did you decide to score it the way you did?" Through their responses, it became obvious that the raters generally agreed upon how to define and identify each sub-category in the rubric. However, according to their accounts of rating individual papers, they tended to remember

weighing certain sub-categories more heavily than others. For example, when rating paper #43, Susan gave it a 13/20 and Jacob gave it a 20/20. Susan weighted the sub-scores of Interpretation and Analysis more heavily in her rating, while Jacob weighed Inference and Explanation more heavily. Since the student did worse on the categories that Susan paid more attention to, she rated the paper lower, while the student did better on the categories that Jacob paid more attention to.

Interestingly, even when raters gave similar scores to papers, they still reported different reasons for their rating. For example, on paper #56, Susan said that the student failed to exhibit critical thinking in interpretation, evaluation, or evidence, but began to show evidence of critical thinking with regard to analysis and disposition. Jacob said that the paper began to show evidence of interpretation, but failed to show any other attributes. Thus, although Susan and Jacob's sub-scores disagreed on paper #56, their overall scores are very similar.

The second question that I asked interviewees was this: "What patterns did you notice in papers that scored low? High?" The raters had ready answers for this question, and their answers shared three common threads. First, Susan and Jacob said that low-scoring papers tended not to display empathy or compassion towards opposing viewpoints. Second, all three raters claimed that low-scoring papers tended to adopt extreme political stances as opposed to moderate stances. Third, all raters explained that high-scoring papers also seemed aware of counterarguments, or multiple viewpoints, and responded to those viewpoints as if writing to a real audience.

The third question the interviewees answered was, "What factors influenced you that aren't specifically mentioned in the rubric?" All the raters named several factors that influenced them, and several factors that did not influence them. Derek and Susan both claimed that their personal political views did not influence them; Derek tried to remain objective and Susan actually found it easier to enjoy a paper with an opposing viewpoint to her own, partly because

she felt it was easier to find holes in an argument that she opposed. Yun's article, "How Raters' and Writers' Perceptions of a Topic Affect the Scoring of Compositions" found that there was no significant correlation between raters' personal opinions and the way they rated papers, although one study is probably not sufficient to prove this.

A few other factors may have influenced the raters, although there isn't enough detail in the interviews to explain what effect they may have had on the scores. Derek and Jacob both mentioned feeling fatigued throughout the rating process. They also both reported that they started out the day rating papers more generously, and became harsher later on. Susan also mentioned the fact that some high-scoring papers were grammatically or stylistically unsophisticated, and it felt strange that the rubric had no way to reward good grammar or good style.

Derek and Susan disagreed about whether using different highlighter colors for each sub-category clarified or confused the rating process. Derek said that the task of highlighting the evidence of each sub-category in each paper was a challenging cognitive task. When I asked what made the task difficult, Derek said that highlighting was difficult because sometimes the writing exhibited evidence of several sub-categories in the same phrase or sentence, so when he was highlighting a sentence, he would ask himself, "Do I highlight this blue or green?" In other words, some sentences weren't clearly in one sub-category or another, so he had to make difficult judgment calls. He got into a groove eventually, but it was still hard. To me, this information is helpful; it means that Derek disliked or found it difficult to separate the sub-categories. In contrast, Susan liked highlighting the different elements because it made the structure obvious. In other words, she could notice patterns easier because of the colors. For her, it was a positive thing.

The fourth question the raters addressed was "How did your understanding of the rubric change as time went on?" Table 5 below shows the definition each rater used for each sub-category for comparison purposes. After the table I add further explanation of how each rater understood the rubric.

Table 5. Rater's sub-score definitions

Rubric Category	Derek	Jacob	Susan
A. Interpretation	The writer cites and interprets outside sources accurately.	The writer uses an outside source.	The writer cites and interprets outside sources.
B. Analysis	*No comment except that this category was "hard to grade."*	The writer analyzes (explains) outside sources.	The writer gives space to counterarguments.
C. Evaluation	The paper spends space and time on counterarguments, takes a moderate stance towards the counterargument, and meta-analyzes their own reasoning.	The writer takes a meta-cognitive step outside of the source to analyze the relationship between the writer's argument and the source they just quoted and analyzed. Jacob called this "the turn."	The writer turns to their own argument again. Susan, like Jacob, calls this "the turn."
D. Evidence	*No comment except that this category was "hard to grade."*	The writer uses logos, or evidence to back up their argument. They usually used evidence after a claim.	The writer uses logos.
E. Disposition	Derek called this a feeling on the disposition of the writer.	The writer empathized with the other person's argument.	The writer's arguments are balanced.

Where did the raters find agreement?

All the raters agreed that it was easy to rate interpretation and disposition. Interpretation seemed factual: did the paper include outside sources or not? And disposition was easy for everyone to rate; they said it was obvious whether the student was empathetic, compassionate, and open-minded or not. The ease of grading these categories is reflected in the similar definitions the raters assigned.

It's also important to note that in analysis, evaluation, and evidence, Susan and Jacob agreed with each other almost 100% in definition, while Derek either didn't express a definition

or had one very different from theirs. This probably explains why Susan and Jacob had similar overall scores, but Derek's overall scores were different enough from Susan's that Jacob had to offer a third rating for seven out of the twenty papers that Susan and Derek graded. In contrast, Derek did only two third ratings for papers graded by Susan and Jacob, and Susan completed four extra ratings for papers graded by Derek and Jacob.

Where did the raters disagree?

Although Jacob asserted that the group came to a communal definition on the rubric (apparently they talked through it after the first few papers they rated), the raters still had differing operating definitions for each sub-category, which suggests that further training before rating was necessary. They seemed to relate or link certain sub-categories differently, and even defined some sub-categories differently. As mentioned above, the biggest differences in definition are in the categories analysis, evaluation, and evidence.

Derek had a harder time with the middle three categories. He explained that it was very difficult to rate analysis, evaluation, and evidence because they were so closely related, and because they required him to make judgments. While thinking about the rubric, Derek wondered whether combining analysis, evaluation, and evidence would make rating easier. Interestingly, Susan thought that evidence was very easy to grade. She simply paid attention to any *logos*, which she defined as evidence, the student used to support their own argument.

Susan observed that interpretation and analysis were very closely related, although that's not obvious in Table 5. It seems that she thought that citing and interpreting outside sources often means that the student is citing and interpreting counterarguments. She seemed to judge the evidence sub-category by the student's inclusion and interpretation of outside citations in support of the student's own argument.

Jacob thought that the relationship between interpretation, analysis, and evaluation seemed similar because they all involve the student's judgment. Jacob identified evaluation and disposition as the most important indicators of critical thinking. Although Derek and Susan did not specifically name which attributes they judged as most important, they both mentioned the overall importance of disposition. It seems that although identifying the relationship between the sub-categories was not part of the initial rating task, the raters naturally formed opinions about which sub-categories they considered most important.

Furthermore, Jacob indicated that the scale (one through four) in each sub-category was a little difficult. He felt like it was easy to rate a paper a "1—Fails to do this" or a "4—Exceeds expectations," but it was much harder to distinguish between "2—begins to do this," and "3—does this adequately." Finally, he said that the analytic rubric was difficult to use because it required him to think deeply about very superficial writing. In other words, he thought it felt forced to search for evidence of deep thinking when it perhaps didn't exist.

In summary, Susan and Jacob agreed on their definitions of the various sub-scores, although in the first interview question they revealed that they weighted these sub-categories differently. Derek disagreed with Susan and Jacob on the definitions of analysis, inference, and explanation. This finding likely explains most of the difference in sub-scores because all raters agreed on the meaning of interpretation and disposition.

Discussion

I began this study with four hypotheses: the first, that raters would use the modified CTAR rubric with high inter-rater reliability estimates. In other words, I hoped that the modified CTAR rubric would be a good way to measure critical thinking in college student writing. In one sense, I proved this hypothesis true, since the raters of the student essays agreed on the overall

scores of the essays. But in another sense, the hypothesis was disproved, because of the low inter-rater reliability for the sub-categories on the rubric. There are at least two possible reasons for negative agreement between raters on the sub-scores of the analytic rubric. First, perhaps either the raters did not receive enough training on the rubric or the rubric wasn't clear enough. It's clear from Table 5 that Derek disagreed on the definitions of at least two of the five rubric categories, although Susan and Jacob agreed on each definition.

Second, it's possible that critical thinking is difficult to break into separate sub-categories, and may be easier to score holistically. In the interviews, the raters seemed to think that critical thinking a complex web of hard-to-measure abilities that are difficult to break into separate small categories of skills. One reason for the high inter-rater reliability on the overall scores could be that experienced raters grasp the essence of critical thinking and subconsciously adjust their sub-score ratings to produce an overall score that they feel confident about. Although we cannot prove that is the case, it suggests the possibility that holistic rating may be more reliable than analytic rating when it comes to critical thinking in writing.

The second hypothesis I tested was that there would be a significant relationship between the scores from the earlier holistic rubric used in the 2015 Hansen et al. study and the scores from the analytic rubric used in this study because I assumed that either the holistic rubric was actually measuring critical thinking, but by a different name, or that critical thinking would correlate with a holistic score. There was no significant relationship between the ratings in the Hansen et al. study and the current study, which could be because the raters in the Hansen et al. study were using a rubric that included additional factors beyond critical thinking (such as grammar, style, etc.). Interestingly, the raters in my study remarked that good critical thinking did not always align with good grammar and style. One rater remarked that she found it difficult

to assign a high score to a paper that exhibited good critical thinking but poor grammar. More research would shed additional light on any correlation between critical thinking, grammar, spelling, and rater bias.

The third and fourth hypotheses I tested were (3) that there would be a significant relationship between analytic scores and ACT and GPA scores and (4) that there would be a significant relationship between essay score and gender. There was no significant relationship between the student's GPA and the paper's overall score. There was also no significant relationship between the student's ACT score and the paper's overall score. There was, however, a significant relationship between the student's gender and the paper's overall score ($p = 0.03$).

While it is outside of the scope of this study to investigate the reason for no relationship between essay score and high school GPA or ACT score, this finding is still interesting. It means that general high school "success" is not necessarily a predictor of critical thinking in first-year writing. Students' GPAs are based on many things, including homework preparation, test scores, attendance, diligence, difficulty of courses, and so on; if critical thinking and writing ability are reflected in student GPAs, they are likely to be represented only minimally. Likewise, ACT scores are based on multiple-choice tests, not fine-grained measures of critical thinking, so it may not be surprising that students might have high ACT scores but not perform well on the paper used in this study. The same kind of critical thinking that was defined in the rubric is likely very hard to detect in a multiple-choice test.

While GPA and ACT scores did not have a significant relationship with the analytic critical thinking rubric scores, gender was a significant variable (female students scored higher on the analytic critical thinking rubric than male students, on average). Hansen et al. also found that women scored higher than men in their assessment, although their finding was not

statistically significant. Several scholars and a national study corroborate the finding that adolescent women write better than men, although their writing might not be better at "critical thinking." In 2011, the NAEP found that female students outperformed male students in both eighth and twelfth grade writing exercises. And Willingham et al. and Lane and Stone's studies reflect similar findings. My study's scope does not include finding out the reason for the disparity, although based on many other studies by educational psychologists and human development experts, it's possible that women's cognitive development progresses faster than that of men or that women take writing assignments more seriously, score higher on elements that raters value (like grammar or spelling), or that raters are more biased towards female students. It appears that if any other factors besides gender influence students' critical thinking ability, those factors may be harder to measure than factors measure by GPA or ACT score. Perhaps more subjective data, such as interviews with the students, could reveal predictors and correlates of successful critical thinking.

A final unexpected finding was that those students who scored highest exhibited a different pattern in the expression of their critical thinking than those students who scored lower. Based on an examination of the color-coding of students' critical thinking abilities by the raters, I determined that the high scorers wove counterarguments, analysis, and explanations throughout each paragraph. They seemed to dance through the critical thinking moves somewhat naturally, rather than mechanically adding a solid paragraph of counterargument at some point in their argument. The low scorers tended to address each facet of critical thinking in chunks; they tended to devote one paragraph to their argument, the next paragraph to the counterargument, the next to an analysis, and so on. Theirs was more like the moves of a memorized line dance. Good for practicing, and maybe still fun, but definitely a beginner's kind of dance.

As a result of the interviews, I gathered observations that I hadn't included in my initial planning. For example, all the raters volunteered suggestions about what they think would help freshman students develop critical thinking skills. They based these suggestions not on the analytic rubric but on the impressions they gathered while reading the papers. The raters noted that students often cite sources that reflect a lack of exposure to valid research on counterarguments, and so they agreed that students need to be exposed to research-based information about people who disagree with them. The raters volunteered that students need to get to know what their audiences think through reading, writing papers from the audience's point of view, or other creative experiences that allow them to experience empathy. Jacob and Derek suggested methods for helping students develop audience awareness. Jacob said that students "need to take time to argue the other side," an idea he gathered from Bean's *Engaging Ideas* book. In Jacob's thinking, perhaps American Heritage students should have a follow-up paper in which they must argue the opposite stance to the one they took in this paper. Jacob has instituted a practice in his WRTG 150 class in which his students take magazine clippings and make collages that represent what their audience wants/needs. He thinks that taking the time to empathize is key to developing critical thinking.

Derek's suggestion for helping students develop critical thinking is that we need to expose them to dissonance, or a wide variety of perspectives, in order to challenge them. According to him, the alternative is for everyone to develop tunnel vision in which they block other perspectives that challenge their own, or simply don't see those perspectives at all. This alternative would be to simply allow everyone's confirmation bias to remain intact and for civil dialogue in our society to cease.

Each rater had some unique suggestions as well. Susan thinks that the students don't

seem to research enough, and need to know more in general about the issue they are writing about, in addition to knowing about their audience. It is possible, she observed, that when they choose their own topics, they are more likely to do in-depth research. Perhaps students need a more obvious genre in American Heritage, she suggested, so they can look at models and be instructed in the structure of the genre. Susan also wonders if students lack confidence in voicing their own opinions. In high school, they are sometimes not allowed to use "I," and perhaps they don't understand that new genres allow the use of "I."

My own conclusions about this project include the importance of giving students opportunities to encounter and consider opinions that are different from their own initial opinions, which this American Heritage assignment attempted to do by asking students to include counterarguments in their writing. However, it seems that, from the papers we read, many first-year college students rely on arguments they've gathered from overheard opinions, and do not know enough about most controversial topics to make an educated judgment call. It's hard to know why this is the case; perhaps age and human development is a factor. Perhaps experience is a factor. Many types of extracurricular activities can help students experience and participate in dialogue—volunteer service, study abroad, in-depth research, and participating in debate teams. Since professors have less control over these extracurricular experiences than they do with what happens in the classroom, we need to focus on what we can offer these students within the context of a writing classroom or American Heritage classroom. My hunch is that we should help students experience the dialogic nature of conversations about controversial topics by returning to one of the oldest rhetoric exercises around: design assignments that ask students to argue the side they don't agree with. This experience will give them opportunities to interpret, analyze, and evaluate their arguments and those of other people.

It is still possible that students will complete the assignment without any personal investment, but for those who do invest in the assignment, they will likely find similarities, nuances, and grey areas between opposing sides. The most ideal situation, in my opinion, is that students will discover that most controversial problems don't have a clear-cut perfect solution, and that compromise, understanding, and compassion support the better creation of solutions.

Limitations and Recommendations for Future Research

Although this study suggested some interesting conclusions, it also had several limitations, and those limitations point to the need for continued research. The first limitation is inherent in looking at only one essay without any contact with the student writers, which we did because of funding and time constraints. In the future, this kind of study would benefit from gathering subjective data that one can gather through interviewing students before, during and after their writing happens. The interviews could discover possible factors that influence students' critical thinking abilities outside of gender, high school GPA, and ACT score.

A future study would ideally study more than one student paper. This study's findings are based on one 900-word paper from each student, not a highly reliable assessment. Consider a student who may have been sick the week before the paper was due, or a student who started a new job. These students may have written beneath their actual abilities on this particular paper, but we have no way of knowing that fact without gathering more data.

A final limitation was the small sample size, n = 30. We were only able to rate 30 student papers because of limited funding for the study. With more funding, I could have paid the raters to review more papers. Also, I could have increased the training time and perhaps have eliminated some of the disparities that arose from different understandings of how to apply the rubric.

APPENDIX A: AMERICAN HERITAGE 100 ESSAY ASSIGNMENT

ASSIGNMENT

This semester, you will write four essays. The first three essays will give you practice constructing the various parts of an academic essay, and the fourth essay will give you another opportunity to practice those skills by constructing a full essay from beginning to end. Each essay will come with specific word limits.

The essays are worth the following point totals:

Essay #1: 10 points

Essay #2: 20 points

Essay #3: 50 points

Essay #4: 70 points

Together, the essay assignments are worth 150 points – as much as the final exam. You should, therefore, take each assignment very seriously. Due dates for each essay are listed in the course syllabus and in this document. Please pay careful attention to these dates, as the penalties for late work are severe. The grading rubrics the TAs will follow for each assignment are available on the American Heritage website.

The essays you will write are not book reports or mere summaries of someone else's position or 5-paragraph descriptive essays of the kind many of us wrote in high school. Instead, they should be essays in which you announce a thoughtful, compelling thesis and construct an effective college-level argument, using appropriate evidence and analysis to defend your position. Each essay should include the Honor Pledge (see the "Policies" section of the syllabus).

Your first three essays will be drawn from one of the topic statements below, each of which touches on one or more key themes of the course. The statements are designed to be controversial – people of good will may agree or disagree with them, sometimes passionately. You should critically evaluate the topic you choose. Your goal is to construct a thoughtful, compelling, insightful argument about the topic.

1. Government action invariably means a loss of individual liberty.

2. When designing a government (or writing a constitution), you should focus more on individual rights than on virtue or on community welfare.

3. An individual can achieve the greatest freedom only when bound to a community.

4. Inequality is a natural condition of human society, not a reason for government intervention.

5. Economic systems work best when producers and consumers set virtue aside.

Specific instructions for each of the four essays can be found below. Please feel free to talk to your TA or professor about any questions at *any* stage of the writing process. Teaching students to develop arguments and communicate them clearly is one of the most important things they can do with their time! Office hours or individual appointments are always available for writing conferences. In addition, be sure to read Gordon Harvey's "Elements of the Academic Essay" (available on the American Heritage website) for an overview of key terms that will be helpful as you write.

ESSAY #3

This assignment is your opportunity to expand upon the work you have already done and to develop a full essay that includes an introduction and thesis, evidence, consideration of one or more counter-arguments, and a conclusion. You will write on the same topic you chose for the first two essays. You should revise your introduction, thesis, and evidence after considering the feedback you received from Essay #2. For this essay, you should add any additional evidence and analysis that is needed to fully develop your argument. You should also consider and respond to at least one counter-argument. The essay should include a thoughtful conclusion. You are limited to NO MORE THAN 900 words.

This assignment is worth 50 points and is due at the beginning of your lab on October 11th or 12th. As with the previous essays, please submit an electronic version to Learning Suite prior to your lab, and you should also bring a hard copy to give directly to your TA. This deadline is firm, and failure to meet it will bring the late penalties described in the syllabus.

Writing this essay will help you learn to...

- Structure an essay effectively. Writing an effective academic essay is not an exercise in stream of consciousness or associative writing, nor is it a mere summary of the arguments of others. You should develop an argument of your own that grows and develops (and does not merely restate the thesis multiple times). As you develop your argument, you should articulate and respond to potential counter-arguments to your main idea. Every good idea has potential counter-arguments, and you should engage with those in the course of your essay. Choose the strongest counter-argument you can think of, not a weak or straw-man objection. If you cannot think of a suitable counter-argument, your thesis is probably not as strong as it could be.

ADVICE FOR ESSAY #3

Once again, Gordon Harvey's definitions can be helpful as you think about your essay. Harvey defines *structure* as

> the sequence of main sections or sub-topics, and the turning points between them. The sections should follow a logical order, and the links in that order should be apparent to the reader (see "stitching"). But it should also be a progressive order—should have a direction of *development* or *complication*, not be simply a list or a series of restatements of the thesis ("Macbeth is ambitious: he's ambitious *here*; and he's ambitious *here*; and he's ambitions *here*, too; thus, Macbeth is ambitious"). And the order should be supple enough to allow the writer to explore the topic, not just hammer home a thesis.

Harvey's advice about the development and complication is especially important. The order of the ideas and evidence matters. Your goal is not merely to restate the thesis over and over (for example, the many ways Macbeth is ambitious), but to construct an argument that progresses or develops as you move from point to point.

In order to receive full credit, you must also engage with a strong counter-argument. This is an opportunity to improve your essay by including a moment of reflection. Our writing guru Gordon Harvey says that *reflection* occurs

> when you pause in your demonstration to reflect on it, to raise or answer a question about it—as when you (1) consider a *counter-argument*—a possible objection, alternative, or problem that a skeptical or resistant reader might raise; (2) *define your terms or assumptions* (what do I mean by this term? or, what am I assuming here?); (3) handle a newly emergent concern (but if this is so, then how can X be?); (4) draw out an *implication* (so what? what might be the wider significance of the argument I have made? what might it lead to if I'm right? or, what does my argument about a single aspect of this suggest about the whole thing? or about the way people live and think?), and (5) consider a possible *explanation* for the phenomenon that has been demonstrated (why might this be so? what might cause or have caused it?); (6) offer a *qualification* or limitation to the case you have made (what you're *not* saying). The first of these reflections can come anywhere in an essay; the second usually comes early; the last four often come late (they're common moves of conclusion).

As Harvey points out, the conclusion is another opportunity for reflection. The conclusion should not include new evidence or a new aspect of your argument, but it should be an opportunity to pause and discuss the implications of the argument you have made. The best conclusions will include these moments of reflection and will not be limited to summarizing or reviewing the points you have already made. (In a short essay like this, a long review of your argument is probably not necessary.)

APPENDIX B: SAXTON ET AL. CTAR RUBRIC

Score	Interpretation	Analysis	Score
6	• Clearly and accurately identifies all of the major viewpoints. • Accurately interprets evidence, statements, graphics, questions, etc. with precision and detail. • Demonstrates confident ability to work with the key concepts and terminology.	• Thoughtfully analyzes all points of view to present a thorough evaluation of similarities and differences. • Accurately identifies important claims, arguments, patterns, and/or assumptions in the evidence. • Consistently demonstrates clear, accurate, detailed and comprehensive ability to organize the information for further examination.	6
5	• Clearly identifies all of the major viewpoints. • Accurately interprets evidence, statements, graphics, questions, etc. • Demonstrates a strong ability to work with the key concepts and terminology.	• Analyzes all points of view to present a thorough evaluation of similarities and differences. • Accurately identifies claims, arguments, patterns, and/or assumptions in the evidence. • Consistently demonstrates an accurate and detailed ability to organize the information for further examination.	5
4	• Identifies not only the major viewpoints, but recognizes some of the nuances of those positions. • Interprets evidence, statements, graphics, questions, etc. • Demonstrates a clear ability to work with the key concepts.	• Analyzes all points of view to present an evaluation of similarities and differences. • Identifies claims, arguments, patterns, and/or assumptions in the evidence. • Demonstrates clear ability to organize the information for further examination.	4
3	• Identifies only the basics of each viewpoint, relying heavily on quotes and failing to articulate points in own words. • Interprets some evidence, statements, graphics, questions, etc. • Demonstrates an uneven or shaky ability to work with the key concepts.	• Analyzes all points of view to present an evaluation of obvious or oversimplified similarities and differences. • Superficially identifies the basic claims, arguments, patterns, and/or assumptions in the evidence. • Demonstrates an adequate ability to organize the information for further examination.	3
2	• Identifies few viewpoints or instead identifies only personal position or point of view. • Offers incorrect or no interpretations of evidence, statements, graphics, questions, etc. • Demonstrates an extremely limited ability to work with the key concepts.	• Presents a superficial analysis of similarities and differences between the various points of view. • Incorrectly identifies claims, arguments, patterns, and/or assumptions in the evidence. • Demonstrates an inadequate ability to organize the information for further examination.	2
1	• Does not identify the viewpoint, but offered a biased position based on previously held beliefs. • Offers no or only biased interpretations of evidence, statements, graphics, questions, information, or the points of view of others. • Demonstrates no ability to work with the key concepts.	• Presents little to no analysis of similarities and differences between the various points of view. • Does not identify claims, arguments, patterns, and/or assumptions in the evidence. • Demonstrates no ability to organize the information for further examination.	1

Score	Evaluation	Inference	Score
6	• Identifies the salient arguments (reasons and claims) from multiple perspectives with a clear explanation of each perspective. • Thoughtfully analyzes and evaluates all major alternative points of view.	• Demonstrates confident ability to apply or extend key concepts to make predictions, drawing inferences, and analyzing implications. • Demonstrates surprising/insightful ability to take concepts further into new territory with broader generalizations and implications.	6
5	• Identifies the salient arguments (reasons and claims) from multiple perspectives. • Offers analyses and evaluations of most alternative points of view.	• Demonstrates a clear ability to apply or extend key concepts to make predictions, drawing inferences, and analyzing implications. • Demonstrates strong ability to take concepts further into new territory with broader generalizations and implications.	5
4	• Identifies relevant arguments (reasons and claims) from multiple perspectives. • Offers analyses and evaluations of alternative points of view.	• Demonstrates an adequate ability to apply or extend key concepts to make predictions, drawing inferences, and analyzing implications. • Demonstrates an adequate ability to take concepts further into new territory with broader generalizations and implications.	4
3	• Superficially identifies some arguments (reasons and claims) from main perspectives. • Superficially evaluates obvious alternative points of view.	• Demonstrates a shaky ability to apply or extend key concepts to make predictions, drawing inferences, and analyzing implications. • Demonstrates an uneven ability to take concepts further into new territory with broader generalizations and implications.	3
2	• Hastily dismisses relevant counter-arguments. • Ignores obvious and important alternative points of view.	• Demonstrates inadequate ability to apply or extend key concepts to make predictions, drawing inferences, and analyzing implications. • Demonstrates a superficial ability to take concepts further into new territory with broader generalizations.	2
1	• Fails to identify relevant counter-arguments. • Ignores all alternative points of view.	• Demonstrates no ability to apply or extend key concepts to make predictions, drawing inferences, and analyzing implications. • Demonstrates no ability to take concepts further into new territory with broader generalizations.	1

Score	Explanation	Disposition	Score
6	• Explicitly integrates key sources to support conclusions that address the question. • Clearly justifies and explains assumptions and reasons with evidence. • Demonstrates warranted, judicious, non-fallacious conclusions by using strong, persuasive support.	• Objectively follows where evidence leads by considering the provided context. • Student demonstrates relativist view of knowledge through the adoption of a consistent point of view with appropriate justification as well as awareness of alternative viewpoints.	6
5	• Integrates multiple sources to support conclusions that address the question. • Justifies and explains some assumptions and reasons with evidence. • Demonstrates warranted, non-fallacious conclusions by using strong support.	• Fair-mindedly follows where evidence leads by considering the provided context. • Student demonstrates relativist view of knowledge through the adoption of a clear point of view with appropriate justification as well as awareness of alternative viewpoints.	5
4	• Utilizes information from several sources, but excludes an important view point. • Justifies and explains reasons with evidence. • Conclusions appear reasonable through use of support.	• Fair-mindedly follows where evidence leads by addressing the provided context. • Student demonstrates an understanding of the existence of multiple perspectives.	4
3	• Correctly references information from few sources, but excludes any sources that support a conflicting view. • Justifies and explains some reasons with evidence. • Conclusions are acceptable, but support is weak.	• Follows where evidence leads, but fails to consider the provided context. • Student demonstrates an understanding of the existence of multiple perspectives, but struggles to evaluate these diverse perspectives.	3
2	• Misuse of information or vague reference of information from the sources. • Seldom justifies or explains reasons with evidence. • Conclusions are limited because support is lacking.	• Defends only with a single perspective and fails to discuss other possible perspectives, especially those salient to the provided context. • Student demonstrates a dualist view of knowledge through a treatment of the issue in terms of right/wrong, black/white, and good/bad.	2
1	• References information from none of the relevant material. • Does explain or explicitly state reasons with evidence. • Argues using fallacious or irrelevant reasons, and unwarranted, unsupported claims.	• Maintains views based on preconceptions and exhibits close-mindedness or hostility to reason. • Student demonstrates a dualist view of knowledge through a treatment of the issue in terms of right/wrong, black/white, and good/bad, and focuses on only one side of the issue.	1

APPENDIX C: THE SOPER RUBRIC

Your Name: _____
Paper #:_____

For each of the five traits below, mark how well you think the writer exhibits the trait in his/her paper. The five traits are interrelated; however, try to judge each trait separately.

A. Interpretation: When citing information from outside sources in support of an argument, the writer interprets that information correctly and its relevance to their argument.
    1. Fails to do this
    2. Begins to do this
    3. Does this adequately
    4. Exceeds expectations

B. Analysis: The writer picks a relevant counterargument and perceives the relationship between ideas in the writer's own argument and in the counter-argument (i.e. they can see similarities and differences between the two).
    a. Fails to do this
    b. Begins to do this
    c. Does this adequately
    d. Exceeds expectations

C. Evaluation: The writer evaluates the credibility of his or her own argument and the credibility of any counter-arguments.
    1. Fails to do this
    2. Begins to do this
    3. Does this adequately
    4. Exceeds expectations

D. Evidence: The writer clearly explains and supports his/her own argument with sufficient evidence and non-fallacious reasoning.
    1. Fails to do this
    2. Begins to do this
    3. Does this adequately
    4. Exceeds expectations

E. Disposition: The writer shows empathy and open-mindedness for other viewpoints, and displays an ethos of diligence, care, reasonableness, and persistence in attempts to think critically.
    1. Fails to do this
    2. Begins to do this
    3. Does this adequately
    4. Exceeds expectations

APPENDIX D: RATING RESULTS

ID	Hansen et al Avg Score	GPA	ACT	MALE	FEMALE	High or Low Scoring
Paper ID 83	2	3.7	26	0	1	L
Paper ID 56	2.33	3.9	34	1	0	H
Paper ID 15	2.5	3.7	32	1	0	L
Paper ID 19	2.5	3.7	28	0	1	L
Paper ID 89	2.5	3.85	26	1	0	L
Paper ID 99	2.5	3.97	27	0	1	L
Paper ID 51	2.5	3.99	26	1	0	H
Paper ID 76	3	4	31	0	1	L
Paper ID 107	3	3.94	32	0	1	L
Paper ID 7	3	3.9	27	0	1	H
Paper ID 57	3.33	3.8	33	0	1	H
Paper ID 69	3.5	4	25	0	1	L
Paper ID 135	3.5	3.9	22	0	1	L
Paper ID 143	3.5	3.75	24	1	0	L
Paper ID 2	3.5	3.4	30	0	1	H
Paper ID 11	3.5	3.8	32	0	1	H
Paper ID 43	3.5	3.8	23	1	0	H
Paper ID 58	3.5	3.8	27	0	1	H
Paper ID 59	3.5	4	27	0	1	H
Paper ID 46	4	4	32	1	0	L
Paper ID 88	4	3.85	28	0	1	L
Paper ID 122	4	3.95	29	0	1	L
Paper ID 144	4	3.9	25	1	0	L
Paper ID 35	4	4	30	0	1	H
Paper ID 37	4	3.9	27	0	1	H
Paper ID 67	4	4	28	0	1	H
Paper ID 5	4.5	3.87	27	0	1	H
Paper ID 52	4.5	3.9	32	0	1	H
Paper ID 65	4.5	3.99	29	1	0	H
Paper ID 142	5	3.9	29	0	1	L

ID	rater1	o1	s11	s12	s13	s14	s15
Paper ID 83	Jacob	14	2	3	3	3	3
Paper ID 56	Susan	7	1	2	1	1	2
Paper ID 15	Susan	9	2	2	1	2	2
Paper ID 19	Susan	9	3	2	2	1	1
Paper ID 89	Susan	8	2	2	1	2	1
Paper ID 99	Jacob	17	3	4	3	3	4
Paper ID 51	Jacob	12	3	2	2	2	3
Paper ID 76	Susan	10	2	2	2	3	1
Paper ID 107	Susan	10	2	2	2	3	1
Paper ID 7	Susan	11	2	2	2	2	2
Paper ID 57	Jacob	20	4	4	4	4	4
Paper ID 69	Susan	19	4	4	3	4	4
Paper ID 135	Jacob	13	2	3	3	2	3
Paper ID 143	Susan	8	2	2	1	2	1
Paper ID 2	Susan	16	4	3	1	4	4
Paper ID 11	Susan	20	4	4	4	4	4
Paper ID 43	Susan	13	3	2	2	3	3
Paper ID 58	Susan	14	3	3	3	3	2
Paper ID 59	Susan	11	2	2	2	3	2
Paper ID 46	Jacob	20	4	4	4	4	4
Paper ID 88	Susan	12	3	2	2	3	2
Paper ID 122	Susan	10	2	2	2	3	1
Paper ID 144	Jacob	10	2	2	2	2	2
Paper ID 35	Jacob	20	4	4	4	4	4
Paper ID 37	Susan	8	2	1	1	2	2
Paper ID 67	Susan	13	3	2	2	3	3
Paper ID 5	Jacob	13	3	2	3	3	2
Paper ID 52	Susan	17	4	4	3	3	3
Paper ID 65	Jacob	9	1	2	2	2	2
Paper ID 142	Susan	8	2	1	1	2	1.5

Key:
o1: Overall score 1
s11: sub-score 1.1

ID	rater2	o2	s21	s22	s23	s24	s25
Paper ID 83	Derek	7	2	1	1	2	1
Paper ID 56	Jacob	5	2	1	1	1	1
Paper ID 15	Jacob	11	3	2	2	2	2
Paper ID 19	Derek	15	3	4	3	2	3
Paper ID 89	Derek	10	3	2	1	3	1
Paper ID 99	Derek	12	2	2	3	2	3
Paper ID 51	Derek	12	3	2	2	3	2
Paper ID 76	Derek	13	3	2	3	3	2
Paper ID 107	Jacob	10	2	2	2	2	2
Paper ID 7	Jacob	12	3	3	3	3	3
Paper ID 57	Derek	20	4	4	4	4	4
Paper ID 69	Jacob	15	3	3	3	3	3
Paper ID 135	Derek	10	2	2	1	3	2
Paper ID 143	Derek	14	3	3	3	2	2
Paper ID 2	Derek	10	2	2	1	2	3
Paper ID 11	Derek	9	2	2	2	1	2
Paper ID 43	Jacob	20	4	4	4	4	4
Paper ID 58	Derek	20	4	4	4	4	4
Paper ID 59	Jacob	18	3	4	4	3	4
Paper ID 46	Derek	12	4	2	1	3	2
Paper ID 88	Jacob	9	2	2	1	2	2
Paper ID 122	Derek	16	4	3	3	3	3
Paper ID 144	Derek	16	3	3	4	3	3
Paper ID 35	Derek	17	4	4	3	3	3
Paper ID 37	Derek	12	2	3	1	3	3
Paper ID 67	Jacob	13	3	2	2	3	3
Paper ID 5	Derek	11	3	2	1	3	2
Paper ID 52	Derek	15	4	3	3	2	3
Paper ID 65	Derek	8	2	1	2	2	1
Paper ID 142	Jacob	11	2	2	2	2	3

ID	rater3	o3	s31	s32	s33	s34	s35
Paper ID 83	Susan	12	2	3	2	3	2
Paper ID 56							
Paper ID 15							
Paper ID 19	Jacob	9	2	2	2	1	2
Paper ID 89							
Paper ID 99	Susan	15	3	3	4	2	3
Paper ID 51							
Paper ID 76							
Paper ID 107							
Paper ID 7							
Paper ID 57							
Paper ID 69							
Paper ID 135							
Paper ID 143	Jacob	9	2	1	2	2	2
Paper ID 2	Jacob	13	2	3	3	2	3
Paper ID 11	Jacob	16	3	4	3	3	3
Paper ID 43	Derek	11	3	2	2	2	2
Paper ID 58	Jacob	13	2	3	3	2	3
Paper ID 59	Derek	17	3	4	3	3	4
Paper ID 46	Susan	9	2	2	1	2	2
Paper ID 88							
Paper ID 122	Jacob	13	3	3	3	2	2
Paper ID 144	Susan	13	2	3	2	3	3
Paper ID 35							
Paper ID 37	Jacob	13	2	3	2	3	3
Paper ID 67							
Paper ID 5							
Paper ID 52							
Paper ID 65							
Paper ID 142							

WORKS CITED

Bean, John C. *Engaging Ideas: The Professor's Guide to Integrating Writing, Critical Thinking, and Active Learning in the Classroom, 2nd ed*. San Francisco: John Wiley & Sons, 2011. Print.

Beaufort, Anne. *College Writing and Beyond: A New Framework for University Writing Instruction*. Logan, UT: Utah State UP, 2007. Print.

Beck, Jacqueline Pinos. "A Longitudinal Study of Critical Thinking Skills in Freshman Nursing Students." MA thesis. Pennsylvania State U, 1999. Ann Arbor: UMI, 1999. Print.

Camp, Heather. "The Psychology of Writing Development—and Its Implications for Assessment." *Assessing Writing* 17 (2012): 92-105.

Haidt, Jacob. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. New York: Vintage Books, 2013. Print.

Hansen, Kristine, Brian Jackson, Brett C. McInelly, and Dennis Eggett. "How Do Dual Credit Students Perform on College Writing Tasks After They Arrive on Campus? Empirical Data from a Large-Scale Study." *Writing Program Administration* 38.2 (Spring 2015): 56-92. Print.

Hunt, Kellogg W. *Grammatical Structures Written at Three Grade Levels*. Champaign, IL: NCTE, 1965. Print.

Karras, Ray W. "An Assembly Plan for Problem-Centered Research Essays." *The History Teacher* 10.1 (1976): 7-19. Print.

Lane, Susanne and Clement A. Stone. "Performance Assessment." *Educational Measurement*. 4th ed. Ed. Robert L. Brennan. Robert L. Brennan. Westport, CT: Praeger, 2006. 387–431. Print.

Lazere, Donald. "Critical Thinking in College English Studies." *ERIC Digest.* Urbana: ERIC Clearinghouse on Reading and Communication Skills, 1987. Print. 2-3.

McGuire, Lauren A. *Improving Student Critical Thinking and Perceptions of Critical Thinking through Direct Instruction in Rhetorical Analysis.* Diss. Capella U, 2010. Ann Arbor: UMI, 2010. Print.

Moffett, James. *Teaching the Universe of Discourse*. Boston: Houghton Mifflin, 1968. Print.

National Assessment of Education Progress. "The Nation's Report Card: Results of the 2009 NAEP High School Transcript Study." 2009. PDF. Web. 20 Nov 2013.

Nussbaum, Michael E. "Using Argumentation Vee Diagrams (AVDs) for Promoting Argument-Counterargument Integration in Reflective Writing." *Journal of Educational Psychology* 100.3 (2008): 549-65. Print.

Pennell, Mike. (2000). "Dialogism in Freshman Writing Classes: Web Projects as Dialogic Knowledge-Making." *NCTE 2000 Conference*. Milwaukee, WI: ERIC.

Perry, William G., Jr. *Forms of Intellectual and Ethical Development in the College Years: A Scheme*. New York: Holt, Rinehart, and Winston, 1970. Print.

Saxton, Emily, Secret Belanger and William Becker. "The Critical Thinking Analytic Rubric (CTAR): Investigating Intra-rater and Inter-rater Reliability of a Scoring Mechanism for Critical Thinking Performance Assessments." *Assessing Writing* 17 (2012): 251-270. Print.

Shafer, Gregory. "Higher Level Thinking, Writing, and Democracy Among Community College Students." *Community College Journal of Research and Practice* 37 (2014): 382-387. Print.

Slomp, David H. "Challenges in Assessing the Development of Writing Ability: Theories, Constructs and Methods." *Assessing Writing* 17 (2012): 81-91. Web.

Strom, Brent. *'The Strawberry Grows Under the Nettle': How an Integrated Performance-Based Approach to the Teaching of Shakespeare at the Secondary Level Affects Critical Thinking Skills as Measured by the California Critical Thinking Skills Test.* Diss. Loyola University of Chicago, 2011. Ann Arbor: UMI, 2011. Print.

Thaxton, Lourene Pike. *In Search of Critical Thinking in Undergraduate Education: A Case Study of a Midwestern University's Center for Teaching Excellence.* Diss. University of Arkansas at Little Rock, 2009. Ann Arbor: UMI, 2010. Print.

Walton, Douglas. *One-Sided Arguments: A Dialectical Analysis of Bias*. Albany: State U of New York P, 1999. Print.

Walton, Douglas. *The Place of Emotion in Argument*. University Park, PA: The Pennsylvania State UP, 1992. Print.

Willingham, Warren W., Nancy S. Cole, Charles Lewis, and Susan Wilson Leung. "Test Performance." *Gender and Fair Assessment*. Eds. Warren W. Willingham and Nancy S. Cole. Mahwah: Lawrence Erlbaum, 1997. 55–126. Print.

Yun, Wang. "How Raters' and Writers' Perceptions of a Topic Affect the Scoring of Compositions." *College Teaching* 51.3 (2003): 115-118. Print.