Research Article

Assessing Writing Constructs: Toward an Expanded View of Inter-Reader Reliability

Valerie Ross, *University of Pennsylvania*
Rodger LeGrand, *Massachusetts Institute of Technology*

Structured Abstract

- **Background:** This study focuses on construct representation and inter-reader agreement and reliability in ePortfolio assessment of 1,315 writing portfolios. These portfolios were submitted by undergraduates enrolled in required writing seminars at the University of Pennsylvania (Penn) in the fall of 2014. Penn is an Ivy League university with a diverse student population, half of whom identify as students of color. Over half of Penn's students are women, 12% are international, and 12% are first-generation college students. The students' portfolios are scored by the instructor and an outside reader drawn from a writing-in-the-disciplines faculty who represent 24 disciplines. The portfolios are the product of a shared curriculum that uses formative assessment and a program-wide multiple-trait rubric. The study contributes to scholarship on the inter-reader reliability and validity of multiple-trait portfolio assessments as well as to recent discussions about reconceptualizing evidence in ePortfolio assessment.

- **Research Questions:** Four questions guided our study:

    1. What levels of interrater agreement and reliability can be achieved when assessing complex writing performances that a) contain several different documents to be assessed; b) use a construct-based, multi-trait rubric; c) are designed

DOI: 10.37514/JWA-J.2017.1.1.09

for formative assessment rather than testing; and d) are rated by a multidisciplinary writing faculty?

2. What can be learned from assessing agreement and reliability of individual traits?
3. How might these measurements contribute to curriculum design, teacher development, and student learning?
4. How might these findings contribute to research on fairness, reliability, and validity; rubrics; and multidisciplinary writing assessment?

- **Literature Review:** There is a long history of empirical work exploring the reliability of scoring highly controlled timed writings, particularly by test measurement specialists. However, until quite recently, there have been few instances of applying empirical assessment techniques to writing portfolios. Developed by writing theorists, writing portfolios contain multiple documents and genres and are produced and assessed under conditions significantly different from those of timed essay measurement. Interrater reliability can be affected by the different approaches to reading texts depending on the background, training, and goals of the rater. While a few writing theorists question the use of rubrics, most quantitatively based scholarship points to their effectiveness for portfolio assessment and calls into question the meaningfulness of single score holistic grading, whether impressionistic or rubric-based. Increasing attention is being paid to multi-trait rubrics, including, in the field of writing portfolio assessment, the use of robust writing constructs based on psychometrics alongside the more conventional cognitive traits assessed in writing studies, and rubrics that can identify areas of opportunity as well as unfairness in relation to the background of the student or the assessor. Scholars in the emergent field of empirical portfolio assessment in writing advocate the use of reliability as a means to identify fairness and validity and to create great opportunities for portfolios to advance student learning and professional development of faculty. They also note that while the writing assessment community has paid attention to the work of test measurement practitioners, the reverse has not been the case, and that

conversations and collaborations between the two communities are long overdue.

- **Methodology:** We used two methods of calculating interrater agreement: absolute and adjacent percentages, and Cohen's Unweighted Kappa, which calculates the extent to which interrater agreement is an effect of chance or expected outcome. For interrater reliability, we used the Pearson product-moment correlation coefficient. We used SPSS to produce all of the calculations in this study.

- **Results:** Interrater agreement and reliability rates of portfolio scores landed in the medium range of statistical significance. Combined absolute and adjacent percentages of interrater reliability were above the 90% range recommended; however, absolute agreement was below the 70% ideal. Furthermore, Cohen's Unweighted Kappa rates were statistically significant but very low, which may be due to "kappa paradox."

- **Discussion:** The study suggests that a formative, rubric-based approach to ePortfolio assessment that uses disciplinarily diverse raters can achieve medium-level rates of interrater agreement and reliability. It raises the question of the extent to which absolute agreement is a desirable or even relevant goal for authentic feedback processes of a complex set of documents, and in which the aim is to advance student learning. At the same time, our findings point to how agreement and reliability measures can significantly contribute to our assessment process, teacher training, and curriculum. Finally, the study highlights potential concerns about construct validity and rater training.

- **Conclusion:** This study contributes to the emergent field of empirical writing portfolio assessment that calls into question the prevailing standard of reliability built upon timed essay measurement rather than the measurement, conditions, and objectives of complex writing performances. It also contributes to recent research on multi-trait and discipline-based portfolio assessment. We point to several directions for further research: conducting "talk aloud" and recorded sessions with raters to obtain qualitative data on areas of disagreement; expanding the number of constructs assessed; increasing the range and granularity of the numeric scoring scale; and investigating traits that are receiving low interrater reliability scores. We also ask whether absolute agreement might be more useful for writing portfolio assessment than reliability and point to the potential “kappa paradox,” borrowed from the field of medicine, which examines interrater reliability in assessment of rare cases. Kappa paradox might be useful in assessing types of portfolios that are less frequently encountered by faculty readers. These, combined with the identification of jagged profiles and student demographics, hold considerable potential for rethinking how to work with and assess students from a range of backgrounds, preparation, and abilities. Finally, our findings contribute to a growing effort to understand the role of rater background, particularly disciplinarity, in shaping writing assessment. The goals of our assessment process are to ensure that we are measuring what we intend to measure, specifically those things that students have an equal chance at achieving and that advance student learning. Our findings suggest that interrater agreement and reliability measures, if thoughtfully approached, will contribute significantly to each of these goals.

Keywords: Cohen's Unweighted Kappa, ePortfolio, formative assessment, interrater agreement, interrater reliability, interreader reliability, kappa paradox, portfolio assessment, rubric traits, trait adjudication, writing analytics, writing assessment, writing in the disciplines

"Given enough time and ways of measurement people can learn to do anything."

--Dorothy Delay, Starling Professor of Violin, Juilliard School

1.0 Background

1.1 Validity and Reliability in Writing Portfolio Assessment

Historically, writing assessment has been troubled by its dependence on subjective judgments (Attali, 2016; Broad, 2016; Cushman, 2016; Elliot, 2016; Poe & Cogan, 2016; Slomp, 2016). A major challenge has been how to reduce or eliminate that subjectivity in order to produce consistent assessments and, conversely, how to embrace that subjectivity as part and parcel of the social construction of knowledge (Huot, 1996). The most recent research in writing studies assessment takes these concerns a step further by calling for fairness alongside consistency as the twin goals of meaningful assessment, with reliability leading the way (AERA, APA, & NCME, 2014; Kelly-Riley et al., 2016). The goal is to develop assessment that is ethical, empirically reliable, and valid, grounded in current theories and practices of writing studies.

Developed by writing studies practitioners, writing portfolio assessment emerged as an alternative to timed essay tests and their emphasis on reliability. Hamp-Lyons (1991), for example, observed, "It is not enough for us to know that judgments are reliable: We must know too that they are appropriate, meaningful, and useful" (p. 2). In contrast, test measurement practitioners have emphasized reliability in an effort to meet institutional demands that call for assessments to be "properly constructed, reliably scored, and economically handled" (White 1985, p. xiv).

As scholars in writing studies have pointed out in their histories of assessment (Behizadeh & Engelhard, 2011; Hamp-Lyons, 1991; Huot, 1996; White et al., 2015; Yancey, 1999), conversations between test measurement practitioners and writing assessment scholars are long overdue. Significant inroads have been made by a few scholars—for example, Elliot, Hamp-Lyons, Huot, and White—whose expertise bridges this divide. Meanwhile, the concepts of reliability and validity continue to pose challenges for writing portfolio assessment, given that writing theorists' concerns range well beyond confirming or predicting the failures and successes of student writers.

As a collection of documents, rather than a single timed response to a common prompt, writing portfolios have proven to be a considerably more effective vehicle not only for assessing writing competence but, more importantly, for advancing our understanding of what, why, and how well our students write. However, unlike test-oriented single-document assessment, portfolios are an unwieldy bundle of documents that do not lend themselves to the kind of checklist-style ratings that are used on timed documents administered under very controlled test-taking conditions using carefully calibrated construct responses and intensively trained raters, and taken by student writers who are often enough well-trained in the same assessment criteria and in how to game it. The timed essay thus enables very high levels of interrater reliability, but calls into question the validity of what is being measured. Students are putting words on the page, to be sure, but what they are doing bears only a faint resemblance to the actual conditions of authentic writing tasks. The timed test tends to be a better gauge of test-takers' socioeconomic backgrounds—for example, those who can afford test-taking coaches and attend top-tier schools—than of students' ability to meet the rhetorical demands of the many writing situations they will encounter throughout their lives.

Having a significant investment in validity—in the meaningfulness and usefulness of what is being measured—writing portfolio theorists have not been uniformly or consistently concerned with the concept of reliability, which measures the consistency of different raters in the scores they give. While influenced by test measurement expertise, writing theorists have also associated reliability with gatekeeping, standardized testing, and teaching to the test, and thus view it as a mixed blessing, if not a Trojan horse (see, for example, Hamps-Lyon, 2002). In recent years, however, reliability is coming into the limelight, poised to make a major contribution to writing assessment, given its potential to illuminate unfairness in terms of, for example, race and gender, or otherwise expose gaps in ecological validity. From this perspective, the concept of reliability promises to generate important new questions as well as put considerable pressure on a range of assumptions and practices in writing assessment. A powerful example is Elliot's (2016) theory of ethics in writing assessment, which is deeply informed by the social justice implications of defining reliability and validity, and argues that we should situate fairness as the boundaries of both. Along similar lines, Slomp (2016), Poe and Cogan (2016), and Cushman (2016) take philosophical and social construct perspectives to explore the historical dangers of considering validity as a boundary for establishing the parameters of fairness in writing assessment. Indeed, most of the recent work in the field frames writing assessment, whether of individuals or of programs, in terms of the ethical considerations that inform fairness and validity (Broad, 2016; Cushman 2016; Elliot, 2016; Poe & Cogan 2016; Slomp, 2016). Elliot (2016) defines "fairness in writing assessment…as the identification of opportunity structures created through maximum construct representation. Constraint of the writing construct is to be tolerated only to the extent to which benefits are realized for the least advantaged" (para. 7).

Whether of a timed essay or a portfolio of widely ranging documents, writing assessment is an unwieldy task rife with variables, from the selection of contents and purpose of the submitted portfolio, to the nature and intensity of training, the locations, sites, styles, criteria of scoring, the time allotted and the time passed between submission and assessment, the credentials, temperaments, working conditions and commitments of writers and evaluators, and the goals of the assessment process itself. The history of the portfolio assessment process is a window into the longstanding effort of writing studies practitioners to take everything possible into account in order to develop multidimensional, mutually reinforcing links between curriculum, assessment policy, and pedagogical practice (Looney, 2011; Shepard, 2000; Torrance, 1998). As such, portfolio assessment bears with it the potential not only to produce more accurate measurements of student achievements, but also to illuminate the path to greater pedagogical effectiveness and fairness, with learning tools and measures designed to meet the needs of students, teachers, and programs.

1.2 The Current Study

This study hopes to contribute to the important research on inter-reader reliability in portfolio assessment as, first, firmly grounded in writing theory and practice; second, construct-based with multiple traits allowing for more granular analysis; and third, multidisciplinary, developed and performed by writing faculty drawn from across the disciplines who teach in our first-year writing program. Based on a larger than typical sample size, our data indicate that the program has developed a solid and replicable curriculum and process that enables reasonably consistent and reliable portfolio assessment using a multiple-trait rubric, a complex collection of documents, and trained raters drawn from across 24 disciplines, rather than a single discipline, as is more customary in writing assessment. Of particular interest is that our findings point to modest but significant reliability in individual trait adjudication, which allows for more granular analysis and identification of potential sites of unfairness or inconsistency in assessing the work of individuals or categories of students. It promises to also contribute to more precise identification of strengths and weaknesses in curriculum, program objectives, faculty training, and in assessment itself.

2.0 Literature Review

2.1 Readers versus Raters

An important but generally overlooked distinction in the quest for reliability in writing portfolio assessment is a key—and arguably discipline-based—distinction between raters and readers. Yancey (1999) observes that writing portfolio assessment is done by readers, not raters. She argues that while raters are trained to replicate scoring practices, readers are "guided rather than directed by anchor papers and scoring guides. . . and value texts and textual features differently." Most importantly, unlike raters, readers expect to differ and to negotiate, a process that contributes to community standards (p. 493). Of course, what Yancey is describing is the way English and writing studies professors are trained to read. Our experience with faculty trained in a wide range of disciplines suggests that they do not all possess these reading habits and are indeed surprised by the lack of agreement and amount of negotiation required to arrive at consensus. One of the questions that arises in a multidisciplinary study of portfolio assessment is, in fact, the extent to which discipline or other factors shape how we read and assess students' writing.

2.2 Rater Backgrounds

Information on rater backgrounds tends to be quite general. Raters tend to be categorized as simple binaries: novice versus experienced, trained versus untrained (see Attali, 2016). Sometimes raters are described in terms of their academic rank: graduate students, lecturers, tenured faculty (Elliot, 2016; Ross et al., 2016a). Occasionally studies mention the disciplinary backgrounds of their raters. For instance, Good (2012) worked with raters from liberal arts, sciences, and education. Vann et al. (1984) surveyed an equal number of instructors from the social sciences, education, biological and agricultural sciences, and physical and mathematical sciences and engineering. Weigle's (1999) study identifies ESL faculty and teaching assistants as raters. The study conducted by Knoch et al. (2007) focuses on raters with backgrounds in English or English as a second language. Further, this study specifies that all of the raters are university graduates. Raters are seldom described in terms of the kinds of disciplinary and professional backgrounds that, based on our experience with portfolio assessment, and scholarship on writing in the disciplines, often substantially inform their assessment decisions (Thaiss & Zawacki, 2006).

In contrast to the paucity of information on raters' backgrounds in assessment studies, the literature exploring rater cognition is vast. Bejar (2012) observes that "rater cognition is concerned with the attributes of the raters that assign scores to student performances, and their mental processes for doing so." Attali (2016) extends this concept to consider how the multidimensionality of rater cognition can influence the quality of writing assessment. Factors such as comprehension of rubric constructs, amount of time allotted to scoring, the number of documents and page count, as well as personal matters, lead raters to develop individualized strategies to facilitate efficient scoring (Attali, 2016).

Scholars agree that the multidimensionality of raters could influence rater reliability (DeRemer, 1998; Freedman & Calfee, 1983; Knoch et al., 2007; Pula & Huot, 1993; Shohamy et al., 1992; Stock & Robinson, 1987; Torrance, 1998; Weigle, 1999; White, 1984; Wiseman 2012; Wolfe, 1997; Wolfe et al., 1998). How a rater approaches a document is "based on world knowledge, beliefs and values, and knowledge of the writing process" (Wolfe et al., 1998, p. 469). In addition, scoring can differ based on "reading skill, background knowledge, or the physical environment in which scoring takes place" (Wolfe, 1997, p. 89). According to Stock and Robinson (1987), rater expectations may be as influential in determining the final score for a document as the quality of writing. Some raters, according to Vaughan (1991), may search for flaws, be influenced by the first impression of the work, focus on two major attributes, scrutinize grammar, or take a personalized approach to reading. Vaughan (1991) notes that readers do not uniformly internalize scoring criteria that they then evenly apply across papers and students.

2.3 Rater Training

Research suggests that rater reliability can be achieved through calibration (Bejar, 2012; Attali, 2016). Rater training is married to rubric comprehension. As White (1984) notes, training sessions help with internalizing rubric guidelines by joining the rubric categories with samples of student writing. Through training, teachers can develop a better understanding of grade boundaries by participating in norming sessions (Attali, 2016). Training sessions have commonly focused on reviewing rubric criteria and how these present in student work (Myers, 1980; White, 1985; Wolfe, 1997). Rubrics are followed but also constrained and inflected by the rater's cognitive, environmental, and temporal circumstances. This complex ecology can prompt flexibility with agreed-upon rules, rubrics, and discourse community goals, enabling the rater to score documents quickly and potentially valuing time over quality of scores and comments (Attali, 2016). Wolfe (1997) observes that "few…training methods explicitly describe the process through which an evaluation is made" (p. 104). Instead, most training emphasizes lining up rubric criterion with student examples "that justify an assigned score" (Wolfe, 1997, p. 104; DeRmer, 1998). The use of the scoring rubric seems to be contingent upon rater cognition. For scorers, rubrics tend to be "fairly abstract documents" (Bejar, 2012, p. 4).

As Wolfe (1997) notes, referencing Pula and Huot (1993), the ability of raters to adapt to a rubric may be determined by how well their experiences and values align with the values informing the scoring rubric. In fact, "the differences among raters may remain even after training, because of the fundamentally different value systems of different graders" (Bejar et al., 2006, p. 58). Rater effectiveness can be improved as a result of participating in scoring process training (Bejar, 2012). While rater training can help reduce differences between rater scores, "other research has shown that the effects of this training may last only a limited time" (Knoch et al., 2007, p. 27).

2.4 Rubrics and Writing Constructs

There is general agreement in the broader educational community that rubrics are a meaningful and useful tool for teaching and assessing writing (Chun, 2002; Ewell, 1991; Hutchings, 1990; Schneider, 2002), with rubrics viewed as increasing assessment transparency and leading to greater fairness and consistency, as well as providing a means for students to advance their own assessment skills (Brough & Pool, 2005; Huber & Hutchings, 2004). As early as the 1950s, research has shown scoring consistency to be very low when raters do not use detailed analytic assessment practices (Dempsey et al., 2009). It is important to note that within the writing studies community, most research on rubrics has focused on holistic essay grading rather than multiple-trait rubric-based assessment of portfolios (Meier et al., 2006), and empirical research methods are rarely used to evaluate the effectiveness of rubrics (Andrade, Du, & Wang, 2008). In general, multiple-trait rubrics are favored over impressionistic, holistic scoring. In her overview of the history of portfolio assessment, Yancey (1999) pointedly asks whether a single holistic score is even appropriate given the complexity of the materials being assessed. Research focusing on interrater reliability in rubric-based assessment has generally supported the use of rubrics (Boix Mancilla et al., 2007; Dahm, & Newell, 2002; Hafner & Hafner, 2003; Jonsson & Svingby, 2007; Newell, Penny, Johnson & Gordon, 2000; Rezaeia & Lovorn, 2010).

Critics of rubric-based assessment argue that it does not take into account the complexities of writing and instead promotes a narrow, formulaic approach to writing (Kohn, 2006; Wilson, 2006). Others suggest that such criticisms are based on poorly designed rubrics or poorly trained raters who approach rubrics as grading tools rather than heuristics (Andrade, 2006; Mansilla et al., 2007). Some writing theorists agree that rubrics should be rooted in local practices to be most effective, with critics pointing to the limitations of one-size-fits-all rubrics that ignore genre and disciplinary distinctions (e.g., Anson et al., 2012).

Construct-based multiple-trait writing portfolio assessment has been gaining traction in recent years (Behizadeh & Engelhard, 2014; Elliot et al., 2016; Hamp-Lyons, 2016; Mansilla et al., 2009; White, Elliot, & Peckham, 2015). Borrowing from the field of psychometrics, construct-based writing assessment takes as its premise that individual writers have a set of attributes that, while not directly assessable, can be identified indirectly and assessed by means of the texts they produce. White et al. (2015) point to four domains: cognitive (e.g., genre, task, audience, writing process, metacognition); interpersonal (e.g., collaboration, social networking, leadership, ethics); intrapersonal (e.g., openness, conscientiousness, extraversion, agreeableness, and stability); and physiologic (e.g., nerve, attention, and vision capacity). Other scholars interested in multidimensional assessment are considering such factors as discipline and affect (e.g., Bryant and Chittum, 2009; Mansilla et al., 2009; Thaiss & Zawacki, 2006).

3.0 Study Context, Sample Description, and Tools

3.1 Study Context

The data are drawn from portfolio assessments of students enrolled in writing courses in Fall 2014 at the University of Pennsylvania (Penn), an Ivy League university in Philadelphia with an undergraduate population that represents 100 countries and all 50 states. Nearly half of Penn’s undergraduates self-identify as students of color. More than half of students are women, 12% are among the first in their families to attend college, and 12% are international (University of Pennsylvania, 2017). The incoming freshmen cohort averages between 2,500 and 2,600 students, along with about 150 to 250 transfer students each academic year. Testing means for the middle 50% of incoming freshmen (University of Pennsylvania, 2016) were as follows: SAT Critical Reading, 690-790; SAT Math, 710-800; SAT Writing, 700-790; ACT Composite, 32-25.

3.2 Overview of the Writing Program

Penn’s Critical Writing Program (CWP), founded in 2003, is an independent writing-in-the-disciplines program. Students from all four undergraduate schools—Arts and Science, Engineering, Wharton, and Nursing—

are required to take a writing seminar to fulfill the writing requirement. Approximately 90% of the freshman class enrolls in the Critical Writing required seminar each academic year; most of the remainder will take it in their sophomore year. The seminars are, with few exceptions, taught by full-time writing faculty drawn from approximately 24 disciplines. The seminars are rooted in the discipline of the instructor teaching them, and organized around a specific scholarly inquiry or debate (e.g., “Are clinical trials effective?” “Does cognitive neuroscience demonstrate that there is no such thing as free will?”) as distinct from the more conventional theme- or anthology-based writing course, which may focus on a topic (e.g., “Science Writing,” “Writing about Writing,” “Nature,” “Intersectionality”) but not on a particular line of inquiry situated within those broad topic areas. Students self-select into seminars, with their decisions based on guidance from the Program’s Directed Self-Placement webpage, advisors, teacher ratings, and word of mouth, along with concerns about schedules and topic interest.

While unusual in the granularity of its scholarly focus, Penn’s curriculum is conventional in requiring weekly readings, writings, and research. Faculty share a common curriculum, a set of assignments that are based on current writing theory and practice, although each seminar is inflected by the individual faculty member’s discipline and line of inquiry. A selection of assignments from this common curriculum provides the content for student portfolios. Preceding the development of portfolio assessment, the curriculum has a deeply rooted self- and peer-assessment culture that is explicitly tied to its formative and summative assessment strategies.

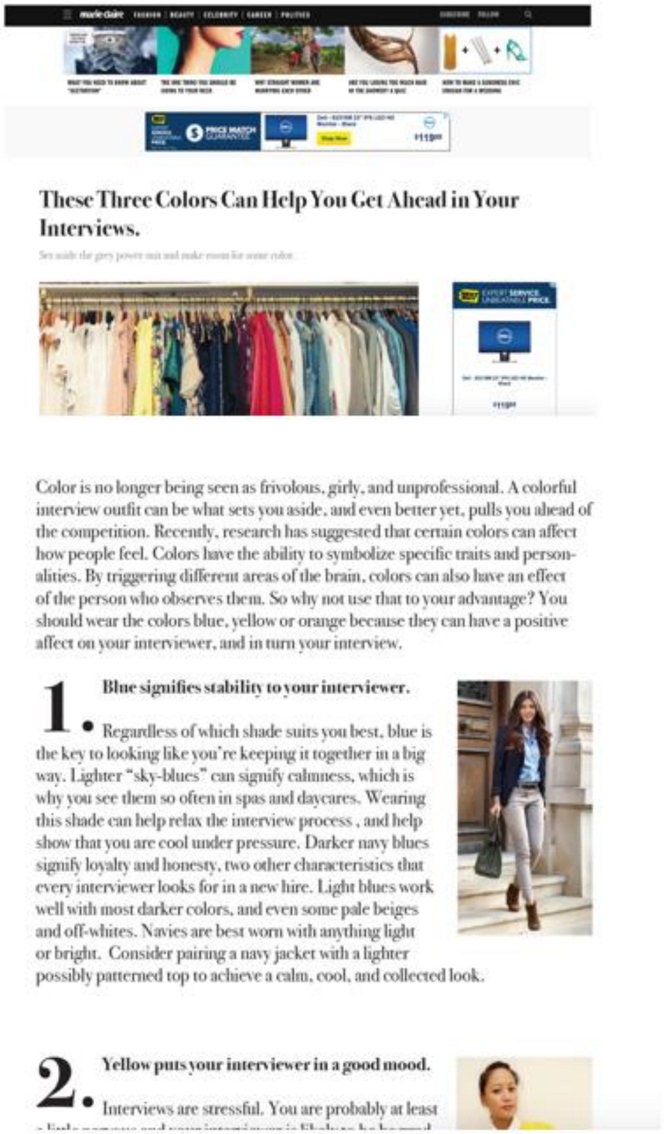
3.3 The ePortfolio

Portfolio assessment was initiated in the program in 2006 concomitant with the development of the shared curriculum. We converted from paper to ePortfolios in Spring 2008. Conducted twice a semester, portfolio evaluation is central to the program’s formative assessment approach, as well as curriculum development, instructor training, and pedagogical practices. The program’s portfolio process and data have also played a significant role institutionally; for example, the 2014 Penn Critical Writing Program’s portfolio assessment process was featured in the 2014 Penn accreditation renewal self-study report for its fairness and consideration of student learning needs (University of Pennsylvania, 2014, p. 102).

Students’ final portfolios average 50 to 80 pages and 17 documents, ranging from a cover letter and resume to peer reviews, self-outlines, timed writing, and drafts. A checklist of contents is provided in Appendix A: 2014 &

2017 Portfolio Contents Details. Students may supplement the required documents with other documents as they wish, although they are obliged to discuss these additions in their final portfolio.

There are two anchor documents that go through extensive research, drafting, and peer review: a digital editorial and an academic literature review. The digital editorial is meant to be visually and rhetorically targeted at an actual publication identified by the student. Here, for example, is an excerpt from a student's editorial for her final portfolio. The student wrote and designed this editorial to simulate actual publication in *marie claire*'s online magazine, which targets an upscale audience of working women.


marie claire
FASHION | BEAUTY | CELEBRITY | CAREER | POLITICS
SUBSCRIBE FOLLOW
PRICE MATCH GUARANTEE
Shop Now
$119
EXPERT SERVICE. UNBEATABLE PRICE.
$119

These Three Colors Can Help You Get Ahead in Your Interviews.

Set aside the grey power suit and make room for some color.

Color is no longer being seen as frivolous, girly, and unprofessional. A colorful interview outfit can be what sets you aside, and even better yet, pulls you ahead of the competition. Recently, research has suggested that certain colors can affect how people feel. Colors have the ability to symbolize specific traits and personalities. By triggering different areas of the brain, colors can also have an effect of the person who observes them. So why not use that to your advantage? You should wear the colors blue, yellow or orange because they can have a positive affect on your interviewer, and in turn your interview.

1. **Blue signifies stability to your interviewer.**

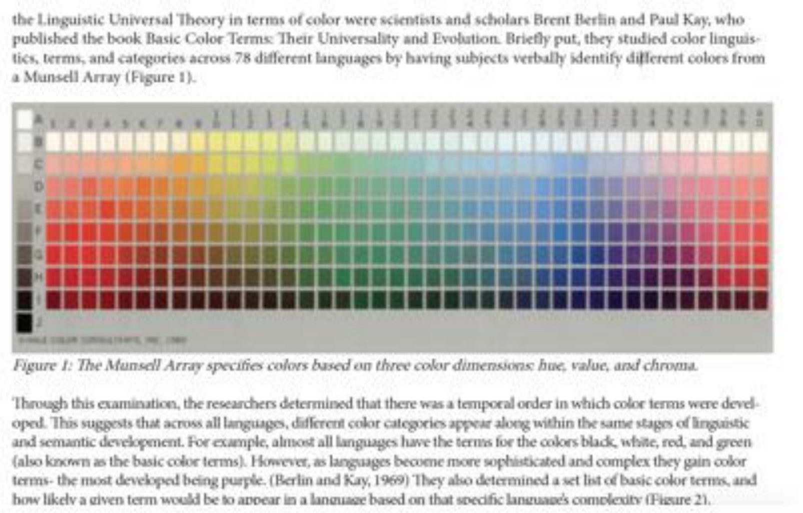
Regardless of which shade suits you best, blue is the key to looking like you're keeping it together in a big way. Lighter "sky-blues" can signify calmness, which is why you see them so often in spas and daycares. Wearing this shade can help relax the interview process , and help show that you are cool under pressure. Darker navy blues signify loyalty and honesty, two other characteristics that every interviewer looks for in a new hire. Light blues work well with most darker colors, and even some pale beiges and off-whites. Navies are best worn with anything light or bright. Consider pairing a navy jacket with a lighter possibly patterned top to achieve a calm, cool, and collected look.

2. **Yellow puts your interviewer in a good mood.**

Interviews are stressful. You are probably at least

Figure 1. Excerpt of student portfolio: Digital editorial.

Unlike the editorial, which is aimed at what may be a very specific public readership (e.g., upscale working women with high school to some college education, 20-35 years old), the literature review is intended for a generalist academic readership of well-educated nonspecialists. Both are based upon the course topic. For instance, this student's literature review was generated by a writing seminar focusing on cognition. The student chose to do a review of literature on color perception and its debates about universalist versus relativist theories of color perception. For the editorial, the student used the knowledge gleaned from research on color perception to persuade readers to pay attention to the colors they use in dressing for job interviews. The student's ingenuity in translating academic expertise into a topic of interest to a particular public readership would likely result in a favorable score in the rubric category of invention, which will be discussed later in this article. Invention is also the category used to assess a student's ability to adapt successfully to different authentic genres.



the Linguistic Universal Theory in terms of color were scientists and scholars Brent Berlin and Paul Kay, who published the book Basic Color Terms: Their Universality and Evolution. Briefly put, they studied color linguistics, terms, and categories across 78 different languages by having subjects verbally identify different colors from a Munsell Array (Figure 1).

Figure 1: The Munsell Array specifies colors based on three color dimensions: hue, value, and chroma.

Through this examination, the researchers determined that there was a temporal order in which color terms were developed. This suggests that across all languages, different color categories appear along within the same stages of linguistic and semantic developement. For example, almost all languages have the terms for the colors black, white, red, and green (also known as the basic color terms). However, as languages become more sophisticated and complex they gain color terms- the most developed being purple. (Berlin and Kay, 1969) They also determined a set list of basic color terms, and how likely a given term would be to appear in a language based on that specific language's complexity (Figure 2).

Figure 2. Excerpt of student portfolio: Literature review.

3.4 The ePortfolio Assessment Process

In addition to weekly rounds of drafting and assessment by the writer, peer, and instructor, students submit midterm and final portfolios for large-scale assessment. The process is relatively straightforward. Students are given a checklist of documents to submit to an online platform in the form of an ePortfolio; in 2014, we used Canvas (See Appendix A for 2014 and 2017 checklists). With the exception of the cover letter, all documents are work already done in class, and the anchor documents, the editorial and literature review, have undergone extensive feedback and polishing. Each student's portfolio is scored and commented upon online by their instructor as well as at least one other outside reader, also an instructor in the program. A rubric is used to guide the assessment process (See Appendix B for 2014 and 2017 rubrics).

The typical arrangement is to put instructors into groups of four to five members who are assigned to read, score, and comment on a sampling of students' portfolios from each of the other's classes. If there are four instructors in a group, each will read one-third of the portfolios of each of the other three instructors; if there are five in the group (we seldom exceed five to keep things lively and focused), then each will read one-fourth of each of the other's portfolios. This allows instructors to get feedback from three or more colleagues from across the disciplines. It exposes them to the quality of work their colleagues' students are achieving, and to the scores and comments that different instructors give to students. They also exchange ideas about teaching the curriculum, working with difficult students, and managing their time. This informal faculty development is especially useful for new or struggling faculty, who get to see how successful instructors score and comment.

Most instructors score about 70 portfolios. Typically, the majority of these are by their own students, with whose work they are already intimate, having given feedback to them along the way. Initial ("unadjudicated") scoring and commenting is done online individually. Thereafter, the instructors meet in person to compare their comments and scores and try to reach consensus whenever there is disparity ("adjudicated" scores). In the early years, this process sometimes became rather acrimonious but was generally sociable. In recent years, perhaps coincident with our program's conversion to a full-time faculty, scoring sessions are usually quite congenial, an opportunity for informal mentoring and a chance to catch up with each other, talk about teaching, and receive welcome feedback on their students and course topics. The atmosphere is relaxed and pleasant. Food is provided. Program administrators circulate to answer questions and to arbitrate when readers have difficulty arriving at consensus. Administrators are required to review any portfolio that has been given an F in any category by either the instructor or the outside reader. Sessions typically take about 1.5 to 2 hours.

Nonadjudicated scores and comments posted by instructors prior to the session, as well as final adjudicated scores, are recorded and saved. Instructors promptly report scores and comments to their students, generally in person. Students are not given access to the unedited remarks made by their instructors and readers. This allows readers to comment quickly and freely rather than having to tinker with their tone or provide extended explanations; those are done conversationally and relayed by individual instructors to their students. On occasion, the comments can be abrupt and harsh (“Worst cover letter I’ve ever seen”) because the instructors know they can clarify when they meet. The instructor is responsible for delivering the scores and comments to students in the most productive and supportive way possible. The overarching goal of the portfolio process is formative—to advance student, instructor, and program learning—and to be fair and consistent in doing so. This also guides how we share feedback and scores with students.

3.5 Raters, Training

In Fall 2014, the Critical Writing Program had 30 full-time instructors from across the disciplines and 10 doctoral candidates from across the disciplines who had been awarded competitive teaching fellowships. Fall 2014 represents the culmination of a 2012 conversion to a full-time NTT faculty (for a discussion of this process and a brief history of the writing program, see Ross et al., 2016b) as a result of a strategic initiative to replace part-time adjuncts with a full-time writing faculty.

CWP faculty and teaching fellows represented the following 24 disciplines in 2014: Ancient History, Anthropology, Archeology, Art History, Cinema Studies, Classical Studies, Cognitive Neuroscience, Communications, Comparative Literature, East Asian Language and Culture, Economics, Education, English, Environmental Studies, Fine Arts, Germanic Languages, History, History and Sociology of the Sciences, Journalism, Molecular Biology, Philosophy, Political Science, Psychology, and Sociology.

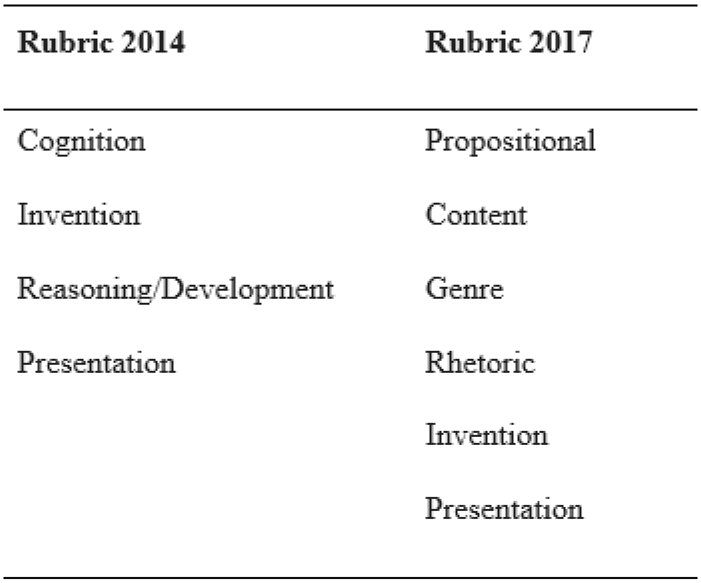
In portfolio scoring and in their classes, instructors rely on the same multi-trait rubric and vocabulary of assessment that our students use in their own self- and peer-assessments. The rubric serves as a guide for formative as well as, at semester’s end, summative assessment. New instructors are extensively trained in our assessment criteria and its symbiotic relationship to all of the activities in our seminars. New faculty engage in norming sessions with the full faculty each fall, and all faculty participate in four rounds of portfolio assessment each year at mid-term and semester’s end. Monthly faculty meetings often focus on curriculum and assessment, and typically result in a discussion of one or more of the rubric traits, sometimes culminating in a vote to change, clarify, or expand the description of a trait. As such, our faculty acquires an intimate understanding of the rubric.

3.6 The Rubric

CWP’s portfolio assessment rubric was introduced in 2006 and indebted particularly to the work of Huot (1996) and White (2005). Since then our rubric has undergone numerous revisions in response to our own research as well as that drawn from a range of disciplines and, most importantly, in response to our faculty's always-evolving views of our pedagogical goals, learning outcomes, and means by which to identify evidence of these in the work our students produce. Below are the rubric categories for 2014 (the focus of this study) and, to bring us up to date as well as show how much the rubric evolves, the 2017 rubric. For detailed explanations of the traits, see Appendix B. Rubrics are scored on a 0 to 4 scale, with 4 representing the highest score and 0 representing the lowest score on a given trait or for the portfolio as a whole.

Table 1

Critical Writing Program Rubrics



Rubric 2014	**Rubric 2017**
Cognition	Propositional
Invention	Content
Reasoning/Development	Genre
Presentation	Rhetoric
	Invention
	Presentation

Our students and writing tutors, like our faculty, are immersed in rubric-based, multi-layered, continuous assessment that functions as a teaching and learning tool. From the weekly feedback students give and receive from instructors and peers, to visits to the Writing Center, students experience a reiterative, multi-layered process of assessment that results in their also being well-trained assessors of writing. This facilitates a shared vocabulary and set of concepts that they can use to communicate with faculty and also with peers, in their writing seminars and as they move along in their academic careers.

In Fall 2016, in part informed by research being done on an NSF grant (National Science Foundation 2015) and in part because of our own and others' research on knowledge transfer and genre, we revised our curriculum and our rubric. While the change in rubric and curriculum are effects of discovery, and also effects of each other—that is, changing the assignments entailed changing the rubric, and vice versa—these changes were more incremental than they appear. While we revised and expanded our traits, the two new traits (genre and propositional knowledge) were already imbedded in earlier descriptions, and the faculty chose to foreground them. The changes reflect a clarification and reorganization of our pedagogical and assessment goals, including a decision to distribute "cognition" across the categories as something that is most readily evaluated in context rather than as a thing in itself. It is easier to see if a student understands how to go about writing an editorial than how to write in general, for example.

Most importantly, the ongoing incremental changes to the rubric represent its provisional nature as we fold in new research, introduce new faculty, and work with a new set of students and peer tutors each year, which all contribute to our ongoing conversations about teaching, learning, and assessment. One concern for us as we embark on this program of assessment research is the extent to which it might discourage changes in assessment criteria or processes for the sake of measuring reliability over time.

3.7 Data Set, Demographics, and Sampling Plan

The data set for this study consists of the non-adjudicated scores of portfolios by the 1,315 students who completed a writing seminar and submitted midterm and final portfolios in Fall 2014. As Elliot et al. (2016) note in their study of portfolio assessment, sample size has been something of an obstacle to portfolio assessment research. Our sample size appears to be about twice as large as what is required for statistical significance (Cohen, 1992; Ellis, 2010).

It's important to underscore the difference between non-adjudicated and adjudicated scores. The non-adjudicated scores that we are focusing on in this article are those that instructors and outside readers give prior to meeting, at which time they discuss any differences in scoring and do their best to achieve consensus across the traits. If an instructor and reader are unable to achieve absolute or adjacent (one point difference) agreement, they will seek the assistance of a third rater, typically a member of the administrative faculty.

In terms of demographics, approximately 85% of the 1,426 students enrolled in the writing seminar in Fall 2014 were freshmen, and 52% were women. While specific information about race and ethnicity is not available for this study, we can provide some demographic context for the Fall 2014 freshman class. Of the 2,350 entering freshmen, 51% were women. Approximately 45% self-identified as white, and just over 40% self-identified as students of color: 19% Asian, 11% Hispanic/Latino, 6% black or African American, 5% as two or more races, and 3 students as American Indian or Alaska native (University of Pennsylvania, 2015).

The sampling plan for the data set includes: SAT Scores = 1,275; Mid-Semester Portfolio Trait Scores = 1,315; Final Portfolio Trait Scores = 1,315; Course Final Grade = 1,315; Term GPA = 1,292.

The specific data in this study is housed in Excel files, which were transferred into SPSS for analysis. The Excel file is organized into 18 columns. The column headings are shown in Table 2 below.

Table 2

Critical Writing Program Fall 2014 Data Table

Column Heading	**Definition**
Instructor	Full name of the course instructor
Reader	Full name of instructor when acting as outside reader
Class Code	Signifies the discipline for the writing seminar topic and the course title. The data set includes writing seminars in 21 disciplines.
Section Number	Unique numerical identifier for each writing seminar
Student Code	(Student Name—anonymized)
Total Average	Represents the total average of instructor and reader rubric scores across the 4 scoring categories identified in our rubric of Cognitive and Heuristic Processes, Invention, Reasoning, and Presentation (see below)
Instructor Grade Cognitive and Heuristic Processes	Instructor’s numerical assessment of student writer's knowledge of writing and rhetorical awareness
Reader Grade Cognitive and Heuristic Processes	Reader’s numerical assessment of student writer's knowledge of writing and rhetorical awareness

Table 2 (continued)

Critical Writing Program Fall 2014 Data Table

Column Heading	Definition
Instructor/ Reader Reliability: Cognitive and Heuristic Processes	The difference between instructor and reader cognitive heuristic processes rubric scores
Instructor Grade Invention	Instructor's numerical assessment of student writer's novelty and persuasiveness for a targeted audience
Reader Grade Invention	Reader’s numerical assessment of student writer's novelty and persuasiveness for a targeted audience
Instructor/ Reader Reliability: Invention	The difference between instructor and reader invention rubric scores
Instructor Grade Reasoning	Instructor’s numerical assessment of student writer's reasonableness and logical coherence
Reader Grade Reasoning	Reader’s numerical assessment of student writer's reasonableness and logical coherence
Instructor/ Reader Reliability: Reasoning	The difference between instructor and reader reasoning rubric scores
Instructor Grade Presentation	Instructor’s numerical assessment of student writer's ability to produce voice, vocabulary, syntax, sentence structure, punctuation, and tone appropriate to the genre and audience
Reader Grade Presentation	Reader’s numerical assessment of student writer's ability to produce voice, vocabulary, syntax, sentence structure, punctuation, and tone appropriate to the genre and audience
Instructor/ Reader Reliability: Presentation	The difference between instructor and reader presentation rubric scores

3.8 Research Method

Fall 2014 nonadjudicated and adjudicated scores and comments were posted by instructors and readers using a University of Pennsylvania software application. Data was exported from the web-based software platform as a .csv file.

To calculate interrater agreement between instructor and reader rubric scores, we populated the following columns from the rubric: Instructor/Reader Reliability: Cognitive and Heuristic Processes; Instructor/Reader Reliability: Invention; Instructor/Reader Reliability: Reasoning; Instructor/Reader Reliability: Presentation.

We used IBM SPSS Statistics V22.0 to produce all of the calculations in this study.

3.9 Calculating Interrater Agreement and Reliability

We used two methods of calculating interrater agreement (IRA): 1) absolute and adjacent percentages, identifying the number of cases in which instructors and their readers agreed upon a particular trait score for a given portfolio, or were off by 1, 2, 3, or 4 points for any of the rubric traits; and 2) Cohen's Unweighted Kappa, which calculates the extent to which interrater agreement is an effect of chance or expected outcome.

Interrater agreement "measures how frequently two or more evaluators assign the exact same rating (e.g., if both give a rating of "4," they are in agreement)" (Graham, Milanowski, & Miller, 2012, p. 5). Interrater agreement is based on a "'criterion referenced' interpretation of the rating scale: there is some level or standard of performance that counts as good or poor" (p.6.)

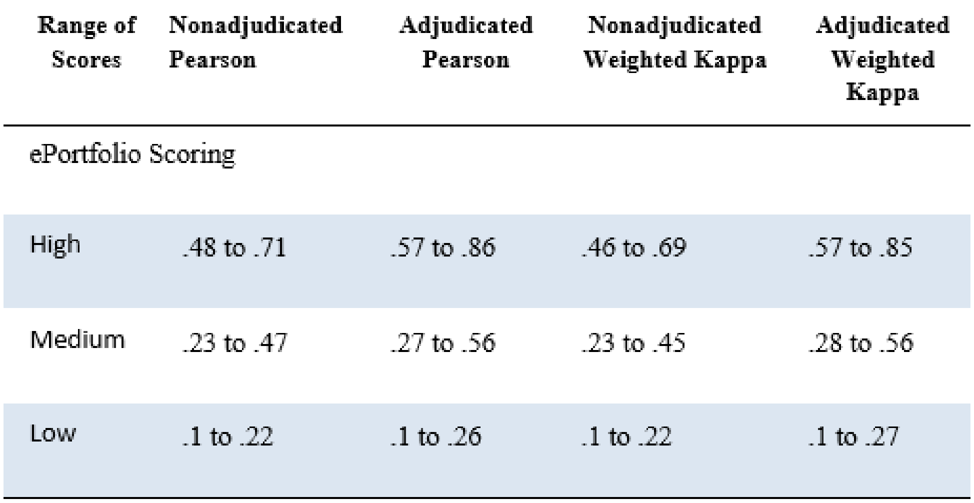
We also measured interrater agreement using Cohen's Unweighted Kappa, which takes into account that agreement may be due to chance alone rather than a thoughtful alignment of two raters' judgments. Kappa is also used when the scale of assessment is narrow, such as in our case, in which raters can only choose from whole numbers, 0 to 4. More reliable ratings are produced by scales that elicit greater granularity, for example, the Likert scale, which allows choices from 0 to 5. The fewer the choices, the greater will be interrater agreement but the less precise the judgments being agreed upon. Kappa is of particular use in bringing more precision to situations like ours.

Kappa helps to grasp how much the observed agreement, calculated by measuring absolute and adjacent rater concordance, differs from chance or "expected agreement." This difference is measured on a -1 to 1 scale, where 1 is perfect agreement, 0 is exactly what would be expected by chance, and negative values indicate less agreement than one would encounter by chance, which is to say, potential systematic disagreement between the observers (Viera & Garat, 2005, p. 361).

For kappa, the benchmark for high agreement is .75 to .80 (Altman, 1991; Fleiss, 1981; Landis & Koch, 1977;). This Kappa scale (Landis & Koch, 1977) is frequently cited.

Table 3

Hypothetical Strength of Interreader Reliability Agreement for ePortfolios



Range of Scores	Nonadjudicated Pearson	Adjudicated Pearson	Nonadjudicated Weighted Kappa	Adjudicated Weighted Kappa
ePortfolio Scoring				
High	.48 to .71	.57 to .86	.46 to .69	.57 to .85
Medium	.23 to .47	.27 to .56	.23 to .45	.28 to .56
Low	.1 to .22	.1 to .26	.1 to .22	.1 to .27

Note: Unless p < .05 on adjudicated scores, the scores should not be used for further analysis. Reprinted from Very like a whale: The assessment of writing programs (p. 123), by White et al, 2015, Boulder: University Press of Colorado. Copyright 2015 by University Press of Colorado.

However, *p* values and confidence intervals are sensitive to sample size, and with a large enough sample size, any kappa above 0 will become statistically significant (Viera & Garat, 2005, p. 362).

4.0 Results

4.1 ePortfolio Scores: Distribution Frequency

Figures 3 and 4 show the distribution of midterm and final portfolio scores in Critical Writing Seminars during Fall 2014.

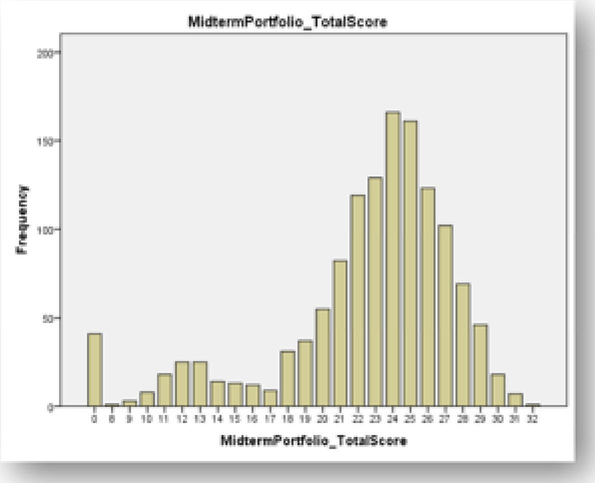


Figure 3. Frequency distribution of midterm portfolio scores Fall 2014.

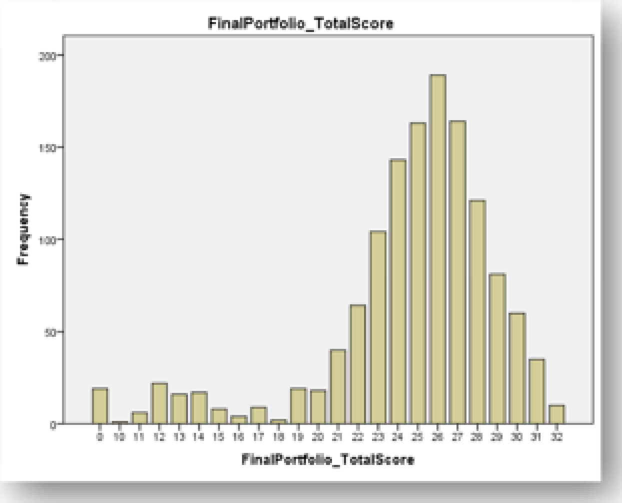


Figure 4. Frequency distribution of final portfolio scores Fall 2014.

4.2 Interrater Agreement: Absolute and Adjacent

The table below shows how often two of our instructor-raters assigned the same rating to each of the rubric traits in Fall 2014.

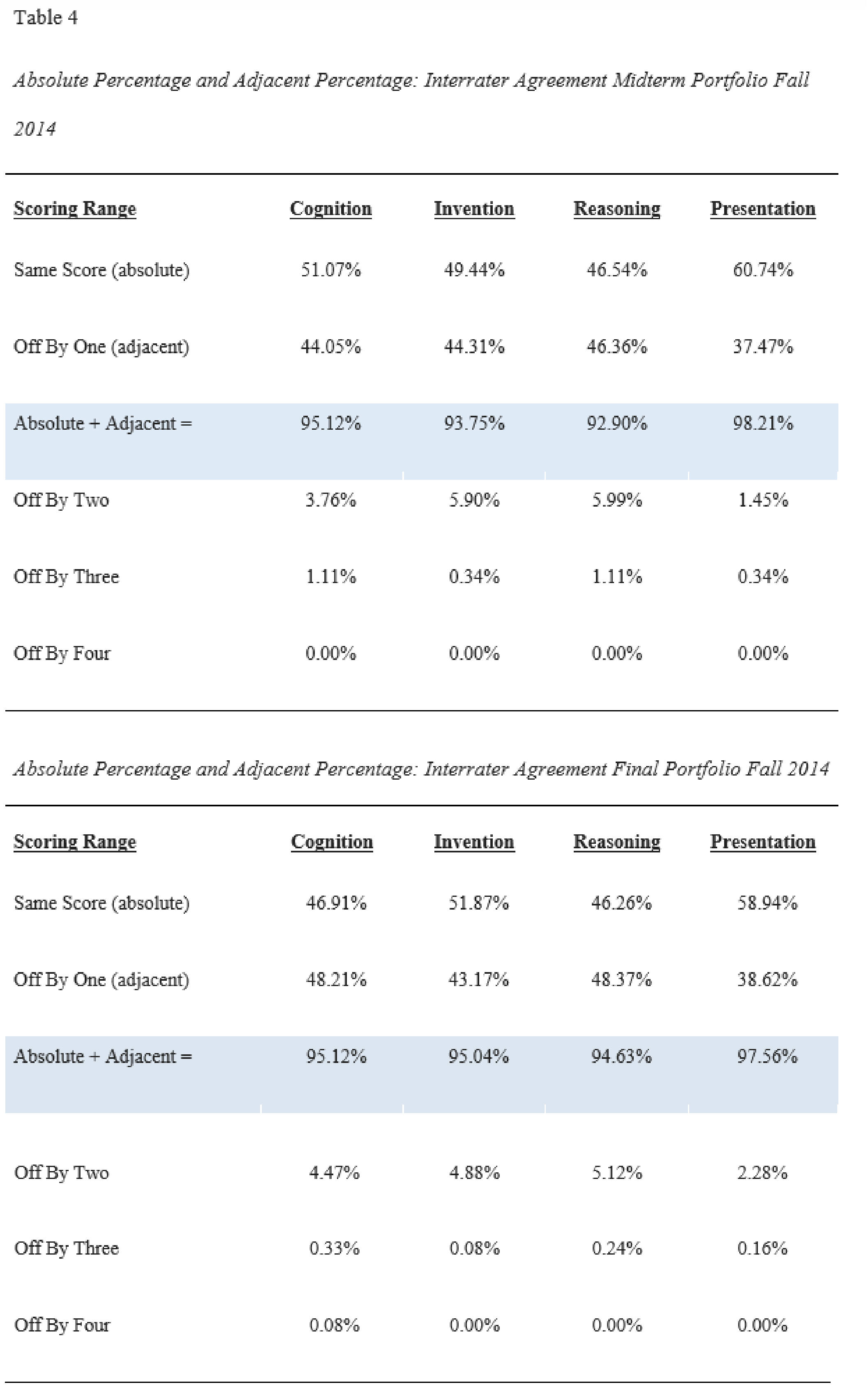


Table 4

Absolute Percentage and Adjacent Percentage: Interrater Agreement Midterm Portfolio Fall 2014

Scoring Range	**Cognition**	**Invention**	**Reasoning**	**Presentation**
Same Score (absolute)	51.07%	49.44%	46.54%	60.74%
Off By One (adjacent)	44.05%	44.31%	46.36%	37.47%
Absolute + Adjacent =	95.12%	93.75%	92.90%	98.21%
Off By Two	3.76%	5.90%	5.99%	1.45%
Off By Three	1.11%	0.34%	1.11%	0.34%
Off By Four	0.00%	0.00%	0.00%	0.00%

Absolute Percentage and Adjacent Percentage: Interrater Agreement Final Portfolio Fall 2014

Scoring Range	**Cognition**	**Invention**	**Reasoning**	**Presentation**
Same Score (absolute)	46.91%	51.87%	46.26%	58.94%
Off By One (adjacent)	48.21%	43.17%	48.37%	38.62%
Absolute + Adjacent =	95.12%	95.04%	94.63%	97.56%
Off By Two	4.47%	4.88%	5.12%	2.28%
Off By Three	0.33%	0.08%	0.24%	0.16%
Off By Four	0.08%	0.00%	0.00%	0.00%

4.3 Interrater Reliability, Unweighted Kappa, and Pearson for Fall 2014

The results for interrater reliability are shown in Table 5. A 0.7 coefficient is typically used as the lowest acceptable correlation for consistency. In section 5.0, we will discuss the limitations of using this standard for writing portfolios.

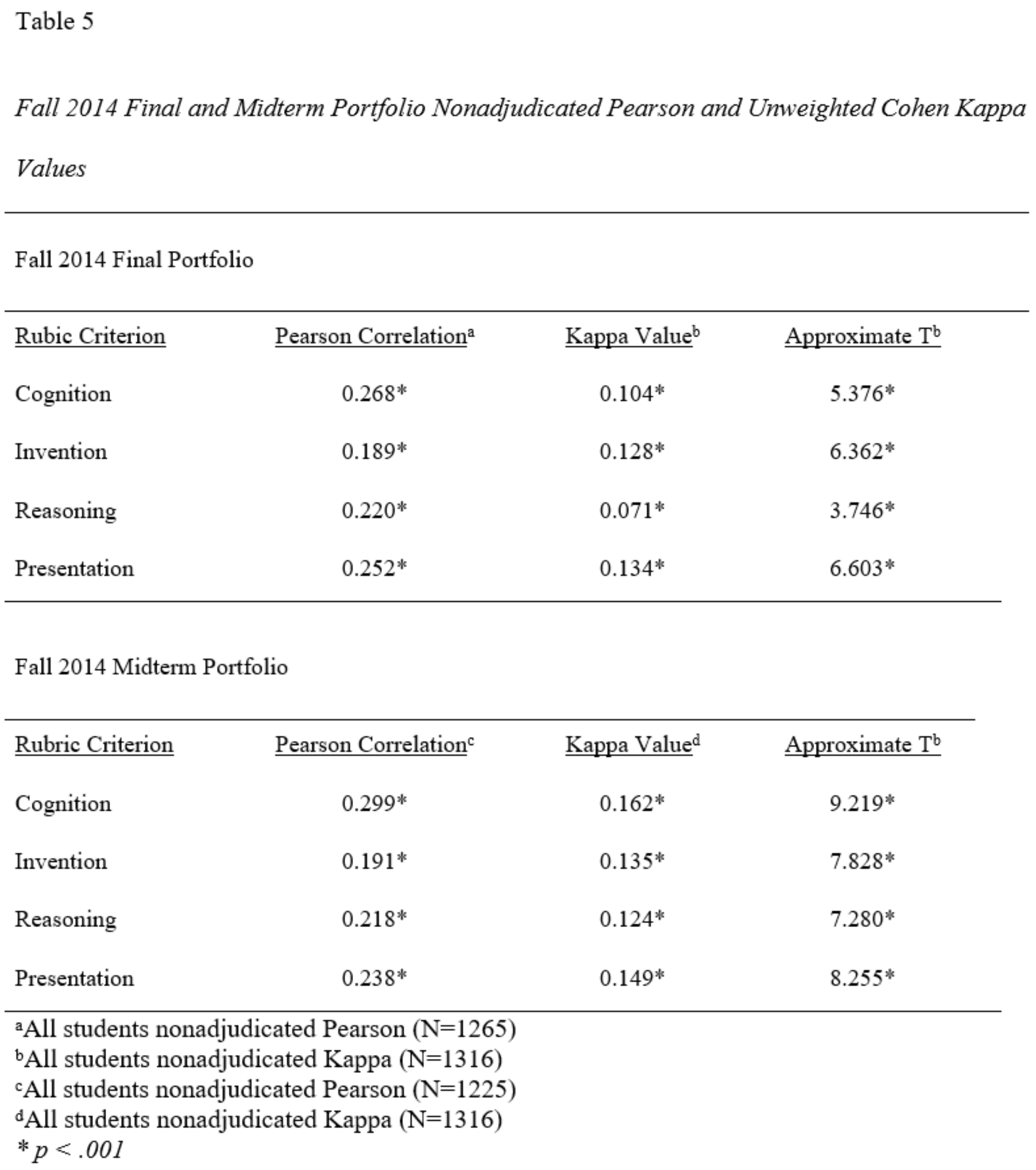


Table 5

Fall 2014 Final and Midterm Portfolio Nonadjudicated Pearson and Unweighted Cohen Kappa Values

Fall 2014 Final Portfolio

Rubric Criterion	Pearson Correlation[a]	Kappa Value[b]	Approximate T[b]
Cognition	0.268*	0.104*	5.376*
Invention	0.189*	0.128*	6.362*
Reasoning	0.220*	0.071*	3.746*
Presentation	0.252*	0.134*	6.603*

Fall 2014 Midterm Portfolio

Rubric Criterion	Pearson Correlation[c]	Kappa Value[d]	Approximate T[b]
Cognition	0.299*	0.162*	9.219*
Invention	0.191*	0.135*	7.828*
Reasoning	0.218*	0.124*	7.280*
Presentation	0.238*	0.149*	8.255*

[a] All students nonadjudicated Pearson (N=1265)
[b] All students nonadjudicated Kappa (N=1316)
[c] All students nonadjudicated Pearson (N=1225)
[d] All students nonadjudicated Kappa (N=1316)
*$p < .001$

5.0 Discussion

5.1 ePortfolio Scores: Distribution Frequency

Accustomed to achieving top grades, Penn students enter our writing seminars expecting to get As from beginning to end. However, as Figure 3: *Frequency Distribution of Midterm Portfolio Scores Fall 2014* and Figure 4: *Frequency Distribution Final Portfolio Scores* demonstrate, writing seminar grades display a typical bell curve for mid-term portfolio and, while grades generally increase (moving to the right on the chart), they still resemble a compressed bell curve at semester's end. This distribution of scores suggests two things: First, despite that these are high-performing students, they enter with different levels of prior knowledge and skills, and exit with improved knowledge and skills. In short, they have things to learn as writers. Second, the improvement demonstrated cannot be attributed to their learning how to write for a particular teacher—a skill Penn students have definitely acquired prior to college—since these scores represent a consensus of their own instructor and an outside reader, whose identity is not shared with them.

5.2 Interrater Agreement, Absolute and Adjacent

The ability of raters to adapt to a rubric is determined in part by how well their experiences and values align with those that inform that rubric (Pula & Huot,1993; Wolfe, 1997). Inoue (2004) points to an ideal of "community-based assessment pedagogy," and later (2007) elaborates on the complexities of value and validation bound up with multidimensional assessment frameworks, complexities that we are truly coming to appreciate even at this early stage of research.

Experts have suggested that absolute and adjacent percentages (i.e., same score or score off by one) should be from 75% to 90% to demonstrate an acceptable level of agreement (Hartmann, 1977; Stemler, 2004). Our raters are above 90% in all criteria traits. However, the ideal is 70% or higher in absolute agreement, although as with all of the calculations we are working with, these ideals are generally derived from studies of simpler, single objects of assessment, such as a timed essay, not the numerous, lengthy, complex documents that populate our portfolios.

There was unusually strong concordance among raters who were in absolute agreement on scores of 2, 3, or 4 in a given rubric category. These account for a large number of cases, a point we will return to in the next section on Cohen Unweighted Kappa analysis.

Overall IRA shows strong absolute-plus-adjacent agreement. Greatest agreement is for the category of presentation, which focuses on the relatively "objective" (rule-bound, acontextual) surface-level issues such as standard edited American English and spelling, as well as formatting (pagination, spacing, fonts) and adhering to citation conventions. We are surprised the agreement isn't even stronger, given how little room there is for interpretation in this category. Note that we include such concerns as wordiness, clarity, and concision in the rubric category of rhetoric, for we maintain that these are issues of style, rooted in the genre, audience, and purpose of the text. Our faculty readers and, by semester's end, our students are knowledgeable of the debates about standard edited English, as well as understand how its "rules" are always evolving. They are aware of differences between prescriptive and descriptive grammar and of discourses on race, class, gender, and nation that problematize the use of standard edited American English. However, faculty and students continue to agree that they need to be able to write for a range of audiences, including those who read high-stakes genres such as applications for jobs, fellowships, and graduate schools, and may be unaware of these important but, as yet, academic debates. Thus. we are puzzled by this relatively low rate of agreement in this category.

5.3 Cohen Unweighted Kappa

We decided to run kappa scores because our scale of assessment was narrow, with raters only being allowed to choose whole numbers from 0 to 4. Kappa is of particular use in bringing more precision to situations like ours to see whether interrater agreement is greater than it might be if instructors scored randomly.

Due to a sufficiently large sample size, our kappa percentages are statistically significant (>0); in other words, our interrater agreement is higher than if we had left it to chance. Nonetheless, these percentages are at the bottom of the scale. We believe that this may be due to what Feinstein and Cichetti (1990) call the "kappa paradox" which occurs when there is a dissymmetry or rarity of the object being rated. Evidence of this possibility is suggested by the fact that the considerable majority of our students are receiving scores in the 3-4 range, with a drop to students earning 2s, and then a significant drop to the infrequent cases of students earning below a 2, particularly by semester's end. As the absolute and adjacent rater agreement rates confirm, there is unusually high concordance of absolute agreement for that range of scores, with agreement dropping noticeably for students who earn below a 2.

Anecdotally, we have long been aware that it is easier for instructors to reach consensus on scores of 4 and 3, less so for scores of 2, and, depending on the situation, much more difficult to agree on scores below 2. In response to this phenomenon, one of our first additions to the portfolio process was to require that all scores below 2 in any rubric category had to be reviewed by administrative faculty. However, even such experienced arbitrators often find themselves seeking a fourth reader because more often than not, the contents of the portfolio are difficult to judge. For example, the student might have a strong command of superficial aspects of writing (e.g., fluent, with sophisticated vocabulary and command of standard edited American English) but demonstrate no ability to adapt to different genres or to develop a substantive proposition. Such students exhibit what Hamps-Lyon has called a “jagged profile,” scoring high in some categories but low in others. Identifying these strengths and gaps in an individual student’s writing is one of the benefits of using multiple-trait rubrics and portfolio assessment, but in these rare instances the issues are more often than not the effects of factors that cannot be addressed by a rubric or conventional pedagogical approaches. Prompted by consideration of the “paradox of kappa,” we realize that in our program, low-scoring portfolios are nearly always, to borrow from the field of medicine, “rare cases,” and need a different approach to assessment. In a culture characterized by high-achieving students, a failing portfolio is typically generated by the sorts of non-cognitive issues that White et al. (2015) point to as essential writing constructs but ones that, at this point, are not addressed by our rubric or the contents of our portfolios: problems with time management, self-confidence, anxiety disorders, learning disabilities, or medical, family, or other emergencies. Until performing this kappa analysis, we hadn’t considered how the complexities of assessing such portfolios were due to their relative rarity, a terrific insight even if we find upon further analysis that the kappa paradox does not account for our low kappa scores.

5.4 Interrater Reliability

There are debates about whether interrater agreement or reliability is the more productive approach to analyzing the scores raters give (Stemler, 2016). Graham et al. (2012) contend that interrater agreement may be of more use in educator evaluations than interrater reliability. Interrater agreement has not received attention from the writing portfolio assessment community, but may be as important as interrater reliability in program-based portfolio assessment, particularly in cases where the portfolio scores affect the student’s final grade or otherwise have a material impact on students. Where interrater reliability measures the relative order and consistency of raters’ judgments—useful for identifying raters who are outliers and need additional training—interrater agreement determines whether raters actually share the same relative understanding of what constitutes excellent or poor performance in the various traits. Such judgments have relatively high stakes for students who, for example, may have to retake our seminars if given 2s or 1s by one of their raters.

It is important to note that the 0.7 standard for reliability that we use in this study is derived from analysis of single, timed, highly calibrated tests, which should be characterized as a highly conservative if not entirely irrelevant gauge for reliability in writing portfolio assessment. Elliot et al. (2016) suggest the following standard for a six-point scale assuming at least $p < .05$: Non-adjudicated low = 0.1 to 0.22, medium = 0.23 to 0.47, and high = 0.48 to 1.00; adjudicated low = 0.1 to 0.26, adjudicated medium = 0.27 to 0.56, adjudicated high = 0.57 to 1.00 (7.1).

As Table 5 shows, the strongest areas of interrater reliability were cognition and presentation. With the exception of "Invention," all of the trait scores landed in the medium level of reliability. The lower reliability for the category of invention merits further analysis. Faculty have expressed difficulties with what is reasonable to demand of students in terms of novelty and originality since students are novices to the discipline, the topic, and the assigned genres.

6.0 Conclusion

This study suggests that a formative, multiple-trait, rubric-based approach to ePortfolio assessment that uses disciplinarily diverse raters can achieve statistically significant rates of interrater reliability and above average interrater agreement percentages. It raises the question of whether higher rates are desirable, given the exigencies of a formative-oriented, program-based writing portfolio assessment, in which differences in reader feedback provide more authentic writing conditions for students. At the same time, agreement and reliability measures point to many avenues of important new research that could lead to substantial improvements in advancing student learning, as well as in writing assessment.

6.1 Research Agenda

Based on our findings in this study, other research we are conducting through the NSF grant supporting this study, and consultation by Norbert Elliot, we have already made three changes to our assessment process: We have expanded our scoring scale from 4 points to 6 points, revised and increased the number of traits being assessed, and added an impressionistic holistic score to help us identify assessment considerations that may not be represented by the rubric but that play a role in how portfolios are judged.

We hope to pursue the following research in the near future:

1. We will collect qualitative data on how readers score and adjudicate differences in scoring, including talk-aloud sessions, in which readers discuss the rationale for the scores they are giving, and recorded sessions of readers meeting to adjudicate scores. This will help us identify strengths, gaps, and obstacles to interrater agreement.
2. We will identify outliers in interrater agreement and reliability, and review their scores in relation to their portfolios in an effort to better understand and address such disparities.
3. We will review the categories of presentation and invention to learn more about why these categories are below average in interrater reliability.
4. Pursuing the kappa paradox, we will follow the suggestion of Viera and Gratar (2005) on how to account for low kappa percentages by distinguishing between agreement on the two sets of findings; for example, we might back out the cases of students receiving scores of 2 or below and consider whether our rubric is failing to provide adequate explanation and traits to allow for stronger rater agreement, or whether such disagreement should be regarded as reasonable and expected.
5. Another question that we hope to pursue is selection: How many documents, and of what sort, are needed for a productive, fair, and reliable assessment?

6.2 Contributions of the Study to the Writing Assessment Community

Our study contributes to the emergent field of reliability and validity studies of writing portfolio assessment in several ways: first, having a statistically meaningful sample size; second, proceeding from formative assessment of a complex writing performance rather than from a test measurement orientation; third, using a robust writing construct featuring a multiple-trait rubric; and fourth, providing assessments by trained, multidisciplinary readers teaching in a first-year writing program.

We hope that our study helps to promote further conversations between test measurement practitioners and writing portfolio theorists and practitioners. Considerable work must be done to build measurements suited to multi-document assessment that are valid, reliable, and fair. Many questions have been raised by this study, including:

1. What are acceptable levels of interrater reliability for writing portfolio assessment as related to score use?
2. Following the lead of some in the field of education, might absolute and adjacent agreement be more important measures than reliability for portfolio assessment?
3. What are the best ways to develop and improve criteria for capturing robust writing constructs?
4. Do the disciplinary backgrounds of readers affect writing assessment?
5. Do other portfolio assessment programs have rare cases that need to be taken into account when engaging in agreement and reliability studies?
6. How do we aim for interrater reliability and agreement while preserving what makes the portfolio assessment process a rich and authentic experience for students and readers alike? At what point, in other words, might striving for reliability interfere with formative assessment?

White (1994) wrote that if writing teachers controlled the assessment process, we would have an expanded version of classroom assessment. It would be time-consuming, expensive, include many different kinds of writing; teachers would participate heavily in the scoring and would provide comments to students; and all of this would provide useful feedback to immediate stakeholders. He cautioned, however, that such a teacher-driven process would resist being reduced to numbers and would not provide the kinds of data that most other interest groups would seek (p.15-16). Representing a program where faculty have for over a decade taken responsibility for writing assessment, we are happy to provide at least a provisional answer to the question of whether one can meaningfully address writing assessment in a way that satisfies the needs of a range of stakeholders (for a discussion of meeting stakeholders' needs, see Ross, 2016). Our early findings suggest that it is not only possible, but desirable, to bring numbers into the picture and thus to provide data that not only other stakeholders, but the faculty, students, and administration of the program, will find meaningful and useful. We are eager to see what this program of portfolio research has to teach us in the years ahead.

Author Biographies

Valerie Ross is the Director of the Critical Writing Program in the Center for Contemporary Writing at the University of Pennsylvania. Her current research focuses on writing in the disciplines, knowledge transfer, curriculum and assessment development, and writing program administration.

Rodger LeGrand is a lecturer in Writing, Rhetoric, and Professional Communication at Massachusetts Institute of Technology. Prior to this appointment he was director of academic administration for the Critical Writing Program at the University of Pennsylvania. Along with teaching, research, and administration, he has five poetry collections, including *Seeds* (2017).

Acknowledgments

This program of research would not have taken place without the encouragement, expertise, and saintly patience of Norbert Elliot; Meg Vezzu's superb editing; and the thoughtful peer reviewers. Thanks also to our dedicated faculty, who read, debate, and rate mountains of portfolios twice a semester; to Brighid Kelly, who organizes the sessions; to all the students who have produced the work we are studying; to our peer tutors, who help make it all happen; to Patrick Wehner, who helped lead the charge back in 2006; and Art Casciato for doing what he could to make this readable. Finally, we would like to dedicate this study to Subarna Banarjee who, in her short time with us, made lasting contributions to our portfolio process. This research was conducted under an Award from the National Science Foundation (1544239) entitled Collaborative Research: The Role of Instructor and Peer Feedback in Improving the Cognitive, Interpersonal, and Intrapersonal Competencies of Student Writers in STEM Courses.

References

Altman, D. (1991). *Practical statistics for medical research* (reprint 1999). Boca Raton, FL: CRC Press.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Andrade, H. L. (2006). The trouble with a narrow view of rubrics. *The English Journal*, *95*(6), 9-9.

Andrade, H. L., Du, Y., & Wang, X. (2008). Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary school students' writing. *Educational Measurement: Issues and Practice*, *27*(2), 3-13.

Anson, C. M., Dannels, D. P., Flash, P., & Gaffney, A. L. H. (2012). Big rubrics and weird genres: The futility of using generic assessment tools across diverse instructional contexts. *Journal of Writing Assessment*, *5*(1).

Attali, Y. (2016). A Comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing, 33(1), 99-115*. Retrieved from http://ltj.sagepub.com/content/33/1/99

Behizadeh, N., & Engelhard, G. (2011). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing Writing*, *16*(3), 189-211.

Behizadeh, N., & Engelhard, G. (2014). Development and validation of a scale to measure perceived authenticity in writing. *Assessing Writing*, *21*, 18-36.

Bejar, I. (2006). *Automated scoring of complex tasks in computer-based testing*. Psychology Press.

Bejar, I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice, 31(3)*, 2-9. Retrieved from http://dx.doi.org/10.1111/j.1745-3992.2012.00238.x

Broad, B. (2016). This is not only a test: Exploring structured ethical blindness in the testing industry. *Journal of Writing Assessment*, *9*(1). Retrieved from http://journalofwritingassessment.org/article.php?article=93

Brough, J. A., & Pool, J. E. (2005). Integrating learning and assessment: The development of an assessment culture. In J. Etim (Ed.) Curriculum integration K-12: Theory and practice (pp. 196-204). Lanham, MD: University Press of America.

Bryant, L. H., & Chittum, J. R. (2013). ePortfolio effectiveness: A(n ill-fated) search for empirical evidence. *International Journal of ePortfolio*, *3*, 189-198. Retrieved from http://www.theijep.com/pdf/IJEP108.pdf

Chun, M. (2002). Looking where the light is better: A review of the literature on assessing higher education quality. *Peer Review*, *4*(2/3), 16-25.

Cushman, E. (2016). Decolonizing validity. *Journal of Writing Assessment*, *9*(1). Retrieved from http://journalofwritingassessment.org/article.php?article=92

Dempsey, M. S., PytlikZillig, L. M., & Bruning, R. H. (2009). Helping preservice teachers learn to assess writing: Practice and feedback in a Web-based environment. *Assessing Writing*, *14*(1), 38-61.

DeRemer, M. L. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing*, *5*(1), 7-29.

Elbow, P., & Yancey, K. B. (1994). On the nature of holistic scoring: An inquiry composed on email. *Assessing Writing*, *1*(1), 91-107.

Elliot, N. (2016). A theory of ethics for writing assessment. *Journal of Writing Assessment*, *9*(1). Retrieved from http://journalofwritingassessment.org/article.php?article=98

Elliot, N., Rudniy, A., Deess, P., Klobucar, A., Collins R., & Sava, S. (2016). ePortfolios: Foundational measurement issues. *Journal of Writing Assessment*, 9(2). Retrieved from http://journalofwritingassessment.org/article.php?article=110

Ewell, P. T. (1991). To capture the ineffable: New forms of assessment in higher education. *Review of Research in Education*, *17*, 75–125.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543-549.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. 2nd ed. New York: John Wiley.

Freedman, S. W., & Calfee, R. C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. Mosenthal, L. Tamor, & S. A. Walmsley (Eds.), *Research on writing: Principles and methods* (pp. 75–98). New York, NY: Longman.

Good, J. (2012). Crossing the measurement and writing assessment divide: The practical implications of interrater reliability in faculty development. *The WAC Journal*, *23*, 19.

Graham, M., Milanowski, A., & Miller, J. (2012). Measuring and promoting interrater agreement of teacher and principal performance ratings. Center for Educator Compensation and Reform. Retrieved from http://es.eric.ed.gov/fulltext/ED532068.pdf

Hafner, J., & Hafner, P. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *Int. J. Sci. Educ.*, *25*(12), 1509-1528.

Hamp-Lyons, L. (1991). *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex Publishing Corporation.

Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing Writing*, *8*(1), 5-16.

Hamp-Lyons, L. (2016). Farewell to holistic scoring. Part Two: Why build a house with only one brick?. *Assessing Writing*, (29), A1-A5.

Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability measures. *Journal of Applied Behavior Analysis*, *10*, 103–116.

Huber, M., & Hutchings, P. (2004). *Integrative learning: Mapping the terrain*. Washington, DC: American Association of Colleges and Universities.

Huot, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication*, *47*(4), 549-566.

Hutchings, P. T. (1990). Learning over time: Portfolio assessment. *American Association of Higher Education Bulletin*, *42*, 6-8.

Inoue, A.B. (2004). Community-based assessment pedagogy. *Assessing Writing*, *9*(3), 208-238. Retrieved from http://dx.doi.org/10.1016/j.asw.2004.12.001.

Inoue, A.B. (2007). A reply to Peter Elbow on a "Community-Based Assessment Pedagogy", *Assessing Writing*, *12*(3). Retrieved from http://tinyurl.com/hah6r5p.

Inoue, A. B. (2015). *Antiracist writing assessment ecologies: Teaching and assessing writing for a socially just future*. South Carolina: Parlor Press.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130-144.

Kelly-Riley, D., Elliot, N., & Rudniy, A. (2016). An empirical framework for ePortfolio assessment. *International Journal of ePortfolio, 6*(2), 95-116.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, *12*(1), 26-43.

Kohn, A. (2006). Speaking my mind: The trouble with rubrics. *English Journal*, *95*(4), 12-15.

Landis, J. R., & Koch, G. G. (1977). A one way components of variance model for categorical data. *Biometrics*, *33*, 671–679.

Looney, J. W. (2011). Integrating formative and summative assessment: Progress toward a seamless system, *OECD Education Working Papers*, No. 58. *OECD Publishing (NJ1).*

Mansilla, V. B., & Duraisingh, E. D. (2007). Targeted assessment of students' interdisciplinary work: An empirically grounded framework proposed. *The Journal of Higher Education*, *78*(2), 215-237.

Mansilla, V. B., Duraisingh, E. D., Wolfe, C. R., & Haynes, C. (2009). Targeted assessment rubric: An empirically grounded rubric for interdisciplinary writing. *The Journal of Higher Education*, *80*(3), 334-353.

Meier, S. L., Rich, B. S., & Cady, J. (2006). Teachers' use of rubrics to score non-traditional tasks: Factors related to discrepancies in scoring. *Assessment in Education*, *13*(01), 69-95.

Myers, M. (1980). *A procedure for writing assessment and holistic scoring*. ERIC Clearinghouse on Reading and Communication Skills, National Institute of Education. Urbana, IL: National Council of Teachers of English. Retrieved from http://files.eric.ed.gov/fulltext/ED193676.pdf

National Science Foundation. (2015). Collaborative Research: The Role of Instructor and Peer Feedback in Improving the Cognitive, Interpersonal, and Intrapersonal Competencies of Student Writers in STEM Courses (Award No. 1544130) Retrieved from https://www.nsf.gov/awardsearch/showAward?AWD_ID=1544130&HistoHistoricalA=false

Newell, J. A., Dahm, K. D., & Newell, H. L. (2002). Rubric development and inter-rater reliability issues in assessing learning outcomes. *Chemical Engineering Education*, *36*(3), 212-215.

Penny, J., Johnson, R. L., & Gordon, B. (2000). Using rating augmentation to expand the scale of an analytic rubric. *The Journal of Experimental Education*, *68*(3), 269-287.

Poe, M., & Cogan, J.A. (2016). Civil rights and writing assessment: Using the disparate impact approach as a fairness methodology to evaluate social impact. *Journal of Writing Assessment*, *9*(1). Retrieved from http://journalofwritingassessment.org/article.php?article=97

Pula, J.J., & Huot, B.A. (1993). A model of background influences on holistic raters. In M.M. Williamson & B.A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 237-265). Cresskill, NJ: Hampton Press.

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, *15*(1), 18-39.

Ross, V., Liberman, M., Ngo, L., & LeGrand, R. (2016a). Weighted log-odds-ratio, informative Dirichlet prior method to compare peer review feedback for top and bottom quartile college students in a first-year writing program. CEUR-WS.org, 1633. Retrieved from http://ceur-ws.org/Vol-1633/ws2-paper4.pdf.

Ross, V., Wehner, P., & LeGrand, R. (2016b). Tap Root: University of Pennsylvania's IWP and the financial crisis of 2008. *College Composition and Communication*, *68*(1), 205-209.

Schneider, C. G. (2002). Can value added assessment raise the level of student accomplishment? *Peer Review*, *4*(2/3), Winter/Spring.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4-14.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters background and training on the reliability of direct writing tests. *Modern Language Journal*, *76*, 27–33.

Slomp, D. (2016). Ethical considerations and writing assessment. *Journal of Writing Assessment*, *9*(1). Retrieved from http://journalofwritingassessment.org/article.php?article=94

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*(4). Retrieved from http://pareonline.net/getvn.asp?v=9&n=4

Stemler, S.E., & Tsai, J. (2016). Best practices in interrater reliability: Three common approaches. In J. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29-49). Thousand Oaks: Sage Publications.

Stock, P. L., & Robinson, J. L. (1987). Taking on testing: Teachers as tester-researchers. *English Education*, *19*(2), 93-121.

Thaiss, C., & Zawacki, T. M. (2006). *Engaged writers and dynamic disciplines: Research on the academic writing life*. Portsmouth, NH: Boynton/Cook Heinemann.

The secrets of Dorothy DeLay's violin teaching methods. (2015, July 13). *The Strad*. Retrieved from http://www.thestrad.com/the-secrets-of-violinist-dorothy-delays-teaching-methods/

Tinsley, H. E. A., & Weiss, D. J. (2000). Interrater reliability and agreement. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 95–124). New York: Academic Press.

Torrance, H. (1998). Learning from research in assessment: A response to writing assessment—raters' elaboration of the rating task. *Assessing Writing*, *5*(1), 31-37.

University of Pennsylvania. (2014). Assessment of Student Learning. *Accreditation and 2014 self-study report*. Philadelphia, PA: Author. Retrieved from the University of Pennsylvania website: https://provost.upenn.edu/initiatives/reaccreditation

University of Pennsylvania. (2015). Common Data Set 2014-2015. Philadelphia, PA: Author, 5. Retrieved from http://www.upenn.edu/ir/commondataset.html

University of Pennsylvania. (2016). Incoming Class Profile. Philadelphia, PA: Author. Retrieved from http://www.admissions.upenn.edu/apply/whatpennlooksfor/incoming-class-profile

University of Pennsylvania. (2017). Introduction to Penn. Philadelphia, PA: Author. Retrieved from http://www.upenn.edu/about/welcome

Vann, R. J., Meyer, D. E., & Lorenz, F. O. (1984). Error gravity: A study of faculty opinion of ESL errors. *TESOL Quarterly*, 427-440.

Vaughan, C. (1992). Holistic assessment: what goes on in the rater's mind? In Hamp-Lyons, L., editor, Assessing second language writing in academic contexts. Norwood, NJ: Ablex, 111-26.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Fam Med*, *37*(5), 360-363.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*(2), 145-178.

White, E. M. (1984). Holisticism. *College Composition and Communications*, *35*, 400-409.

White, E. M. (1985). *Teaching and assessing writing: Understanding, evaluating and improving student performance.* San Francisco, CA: Jossey-Bass.

White, E. M. (1994). Issues and problems in writing assessment. *Assessing Writing*, *1*(1), 11-27.

White, E. M. (2005). The scoring of writing portfolios: Phase 2. *College Composition and Communication*, *56*(4), 581-600. Retrieved from http://www.jstor.org/stable/30037887

White, E. M., Elliot, N., & Peckham, I. (2015). *Very like a whale*. Boulder, Colorado: University Press of Colorado.

Wiggins, G. (1994). The constant danger of sacrificing validity to reliability: Making writing assessment serve writers. *Assessing Writing*, *1*(1), 129-139.

Wilson, M. (2006). *Rethinking rubrics in writing assessment*. Portsmouth, NH: Heinemann.

Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing*, *17*(3), 150-173.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing*, *4*(1), 83-106.

Wolfe, E. W., Kao, C. W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, *15*(4), 465-492.

Yancey, K. (1999). Looking back as we look forward: Historicizing writing assessment. *College Composition and Communication*, *50*(3), 483-503. doi:1. Retrieved from http://www.jstor.org/stable/358862 doi:1

Appendix A
2014 & 2017 Portfolio Contents Details

Final Portfolio Contents: 2014	**Final Portfolio Contents: 2017**
• Cover letter • Two or more drafts of the research essay with post outlines • A classmate's peer review of the writer's research essay draft • The writer's peer review of another student's research essay draft • Draft of the complex synthesis (author and sources) with post outline • Draft of integrated synthesis (keyword) with post outline • Final grammar check • Two 30-minute timed in-class essays • 30 minute diagnostic essay from the beginning of the semester • Baseline document and post outline • Optional materials to support cover letter	• Statement of academic integrity • Table of contents *Publications* • Cover letter • Resume • Final draft of digital editorial • Final draft of literature review *Supporting Materials* • One-on-one peer review of a classmate's literature review and a copy of the document reviewed • On demand (timed) writing 2 • Pre outline and post outline of justificatory genre • Early draft of digital editorial • Classmates' peer reviews of writer's editorial • The writer's multiple reviews of other students' editorials • Revision plan for editorial • Pre and post outline of literature review • Early and midterm draft of literature review • Classmates' peer reviews of writer's literature review • The writer's multiple reviews of peers' literature reviews • Revision plan for literature review • Final grammar check

Appendix B
2014 & 2017 Rubric Details

2014 Rubric

Cognition/ Metacognition: Knowledge of Writing	· Recognizes the purpose of the assignment · Conceives of a procedure for fulfilling it · Perceives the problem(s) to be solved in the assignment · Follows directions through all stages of the assignment · Able to detect flaws in reasoning in one's own or other's reasoning (outlines and peer reviews) · Able to identify and evaluate (in plans, outlines, peer reviews, cover letters, other artifacts): rhetorical strategies Audience Purpose Genre plan/arrangement complex synthesis Presentation

Invention: Idea/Audience (test of novelty, creativity, persuasion)	· selection of an appropriate and engaging subject within the topic · ability to select and work successfully within a genre · selection of an appropriate proposition and reasons to support it, attuned to the audience and purpose · selection of the appropriate amount and type of evidence and materials to support the proposition, attuned to the audience and purpose · arrangement and style attuned to the audience, purpose, and genre, including ability to evaluate the strength of reasons and evidence · identification of shared premises to enable an effective introduction and conclusion · ability to grasp feedback or detect problems with invention and revise accordingly · ability to vary voice and style to accommodate different audiences and genres
Reasoning: Development/ Coherence (test of reasonableness)	· creation or selection of an appropriately justificatory or explanatory proposition · creation or selection of reasons that directly support the proposition · selection of evidence that confirms, illuminates or otherwise develops the reasons · ability to test argument through strategies of counterargument · demonstration of logical coherence: all reasons support the proposition, all evidence supports the reasons, and to the extent possible, reasons do not contradict each other · demonstration of semantic coherence: sentences and paragraphs stick together

Presentation	· Control of vocabulary, syntax, sentence structure, punctuation, tone · Ability to integrate rhetorical strategies and sources so that they create a consistency of style appropriate to the genre and audience · Demonstrated ability to proofread and polish work for an outside reader · Creation and use of grammar checklist to identify context and patterns of error in mechanics and usage, as well as to correct them · Appropriate formatting, citation, documentation of sources

2017 Rubric

Propositional Content:

4 (Distinguished Pass): The writer produces a logically coherent, factual, knowledgeable explanation or argument to demonstrate or prove the proposition (hypothesis, thesis, claim, inquiry). The exceptional writer has a proposition that is succinct, knowledgeable, and appropriately scaled to the genre, field, and task at hand. The writer's propositional *content* exhibits an exceptionally strong understanding of the subject matter and an *internally coherent* logical framework that draws reasonable inferences and conclusions from relevant sources or ideas chosen to directly advance the proposition.

3 (Mid-level Pass): The points made are reasonable, and the author has a solid understanding of the subject matter, but there are some areas where it is not as logically coherent or carefully reasoned as one would find in a distinguished pass. The proposition is solid but may be modestly off the mark in terms of genre, field, or scale. Additionally, there may be minor problems with inaccuracies, gaps in knowledge and thus minor problems in inference or conclusions based on these or on some logical incoherence in the text. Overall, however, the author demonstrates clear understanding of how to formulate an acceptable proposition and to support it with reasons, evidence, and factual, truthful, accurate, relevant information and evidence. There may be minor problems with logical coherence or validity, but overall the logical structure is recognizable, coherent, and sufficiently developed, and the reasons and evidence do relate to a reasonably well-scaled proposition.

2 (Pass): The reasoning is not very precise or discriminating. There may be some mistakes that the author makes when describing the subject matter, showing that they understand some of the content, but have gaps in their knowledge. The proposition may be rooted in reasonable ideas but be too broad or narrow to be effective. Alternatively, there may be a reasonable proposition but the reasons or evidence are not appropriately chosen or sufficiently developed to render the argument or explanation internally coherent. The reasons and evidence may also stray from supporting the proposition and appear tangential rather than logically structured. The reader is able to move through the text and follow it, but there may be some problems with the logical framework or the accuracy or relevance of the information provided to advance it.

1 (Fail): Substantial errors in presenting a logical framework of accurate, factual or relevant propositional content or logical framework. This may be reflected in an unclear, undeveloped, or improperly scaled proposition or one that is inappropriate to the genre and field. Other issues may include significant logical contradictions or an otherwise weak logical framework, faulty inferences or conclusions, significant gaps in knowledge or factual errors.

Invention:

4 (Distinguished Pass): A distinguished pass in invention, in which the author demonstrates novelty and significance of the work, can take many different forms. The writer may provide a new idea or an original solution to an existing problem, or may bring a new perspective. Invention can also be shown through the selection of reasons and evidence, background information, or a particular method; invention can also be demonstrated through identifying new premises or unlinking a set of unquestioned premises. Invention in scholarly work often entails showing how one’s work adds to the current scholarship in the field by filling in a current gap in knowledge, solving a new problem or finding a better solution to a longstanding problem; answering a root question; selecting and synthesizing appropriate sources and examples to demonstrate the work’s impact and relevance. Invention may be demonstrated by a novel proposition or novel connections made in the synthesis of materials. The inventive writer may also find novel ways of engaging readers, from visuals or formatting to arrangements or style, while honoring the constraints and spirit of the genre and the purpose of the work itself.

3 (Mid-level Pass): Engaging proposition that has relevance within the field. However, not all aspects of the material are equally well-treated, so that the significance of some aspects of the text is not as clear as others. For instance, some evidence or reasons may be familiar and predictable. The author may not fully set up the context of the topic, so the audience cannot fully appreciate its relevance or contribution to the field. Or, the writer may do a good job of synthesizing sources and making connections but some of the connections made may feel predictable rather than novel insights. Overall, the writer is able to support a relevant proposition and establish relevance to the field but the reader may not feel consistently convinced of the work’s originality of contribution or style.

2 (Pass): The text will be on a topic that relates to the course, but is overly generic and familiar. The author has not effectively established the relevance of their work or engaged the reader. The essay will have reasons and evidence, but these will be formulaic and will not provide any deeper insights into the field of study than what the reader is already familiar with. The author does not highlight novel connections between research or demonstrate a new approach to a problem. The writer is able to grasp the basics of writing, but approaches it in a manner that is formulaic rather than strategic, and lacks significance in the field.

1 (Fail): Insufficient understanding of how to come up with an appropriate topic, proposition, or evidence. Inability to grasp how to produce a work that is relevant and significant to the field.

Rhetoric:

4 (Distinguished Pass): The author is masterful in tailoring their content to their particular audience, demonstrating a real awareness of what the reader knows, believes, and values, as well as the expectations and motives a reader brings to the work. The writer is thoughtful, attentive, respectful of the reader, and strives to inform as well as shape the reader's attitude. The level at which information is presented matches that of the reader, and strategies such as the use of examples or figures to illustrate a point are pitched precisely to engage and persuade the reader. The selection of reasons and evidence demonstrates a writer in tune with the target audience(s), as does the use of voice and tone. Organizational and other signposts (boosters, hedgers, temporal signals) show strong rhetorical awareness, as does the arrangement of the reasons, the content, and the consistent style of the text. Overall, careful attention has been paid to the most effective way to convey and supplement propositional content, resulting in a successful transfer of ideas to the reader.

3 (Mid-level Pass): The author clearly demonstrates an effort to take the reader into consideration; however, in a few areas the writer may be less successful in translating this into a rhetorically aware text. For example, the author may not have a precise enough grasp of the target readers' needs and prior knowledge. The premises might not be quite attuned to the target audience, or some specialized language might not have been sufficiently defined (or in turn may be too extensively defined). The author may have chosen premises, reasons or examples that were logically valid but perhaps not as persuasive and geared toward the target reader as they might have been. The voice, tone, and style are suitable to the audience and in some aspects of the text demonstrate a strong awareness of the reader. Overall, the writer does have the reader in mind but here and there exhibits an inconsistency or lapse in awareness.

2 (Pass): The author has demonstrated some ability to consider their audience when writing; however, there are consistent indications that the author has focused on internal logic and subject matter to the point of neglecting how best to convey and tailor these to the reader. For example, the premises may be legitimate but not necessarily attuned to the target audience, or the writer may not give consideration to how best to organize the material to help the reader grasp it. There may be an excess of undefined or unnecessary jargon, or a tone that is not appropriate to the genre. While the reader is able to follow the writer's line of reasoning, there may be areas where the reasons are ineffective due to a lack of background information or explanation to illustrate the point. The writer grasps the basics of writing, but is unable to fully persuade or inform the reader of his or her ideas or, in short, to keep the reader's needs in mind.

1 (Fail): Does not demonstrate an awareness of writing to an actual audience and is unable to tailor writing to target readers. The author is not effective at taking the knowledge that they have and communicating it to the reader. He or she does not use good strategies for selecting or organizing their reasons. No consideration is given to the best way to present the content so that the reader is able to follow and engage with the text.

Genre:

4 (Distinguished Pass): The writer demonstrates a deep understanding of the genre, which means not only adhering to the formal features of the genre (for example, titles, subheadings, particular kinds of content and language) but also how these formal features are connected to the social purposes of the genre and its relationship to readers (consider, for example, the different social purposes and relationships to readers one finds in comparing the title of a newspaper story, a novel, an article in a medical journal). Understanding the formal features as making possible certain kinds of social actions or relationships with readers, genre knowledge entails understanding what readers expect from the genre, what reader/writer relationships the genre creates and generates, how, when, where and why to use one genre rather than another, and what the motives are of the genre's author and readers--how, in short, one's readers put to use each of these formal features. A distinguished writer would be aware of which content is obligatory, and the extent to which the genre places responsibility on the writer to explain and predict for the reader, or on readers to do their own predictions and explanations

3 (Mid-level Pass): The writer generally demonstrates solid understanding of the formal features of the genre but does not appear to fully grasp how these features are tied to fulfilling the purpose and social actions of the genre. Thus, for example, the author may provide a conclusion, in keeping with the formal features of the genre, but the author does not understand the social purpose of the conclusion in that particular genre (for example, in some genres the conclusion is a recapitulation of the text; in others, it describes and demands a call to action; in still others, it points to challenges, implications, or new developments). A writer who recognizes the formal features but does not fully understand the social features of the genre may thus be able to recognize and fulfill the social/purposeful expectations of some of the formal features, but not others; for example, unable to distinguish one type of conclusion from another, or why such distinctions are essential to fully realizing a genre's demands.

2 (Pass): The author is able to identify and execute most of the formal features of the genre, but does not demonstrate an understanding of its purpose and social features. Instead, one feels that the writer is at the stage of filling a form and thus grasping the work itself as formulaic.

1 (Fail): Insufficient understanding of the genre and inability to generate a piece of writing that resembles the formal features of the genre.

Presentation:

4 (Distinguished Pass): The writer submits a polished manuscript, which is to say carefully organized, proofread, properly formatted, cited, nearly or entirely error-free and impressive in its attention to presentation in all aspects. Presentation focuses strictly on surface features, from grammar/mechanics to pagination and aesthetics.

3 (Mid-level Pass): The writer submits a generally clean manuscript, though there may be a few errors in grammar/usage, or a few signs of carelessness with such things as formatting, ordering of documents, inconsistent citation practices. These errors, however, do not distract significantly from the reading experience.

2 (Pass): The writer has some issues with grammar/usage but demonstrates basic competence with standard edited American English. There may also be some issues with inclusion or ordering of documents, formatting, or citation practices that are sufficiently distracting to call attention to problems with the quality of the presentation, but not so distracting that they interfere with the ability to read and assess the text.

1 (Fail): Significant problems with presentation-—for example, issues with language proficiency, grammar/usage, missing documents—that prevent the reader from being able to understand or assess the text.