CHAPTER 2.

VALIDITY INQUIRY OF RACE AND SHARED EVALUATION PRACTICES IN A LARGE-SCALE, UNIVERSITY-WIDE WRITING PORTFOLIO ASSESSMENT

Diane Kelly-Riley
Washington State University

This article examines the intersections of students' race with the evaluation of their writing abilities in a locally-developed, context-rich, university-wide, junior-level writing portfolio assessment that relies on faculty articulation of standards and shared evaluation practices. This study employs sequential regression analysis to identify how faculty raters operationalize their definition of good writing within this university-wide writing portfolio assessment, and, in particular, whether students' race accounts for any of the variability in faculty's assessment of student writing. The findings suggest that there is a difference in student performance by race, but that student race does not contribute to faculty's assessment of students' writing in this setting. However, the findings also suggest that faculty employ a limited set of the criteria published by the writing assessment program, and faculty use non-programmatic criteria—including perceived demographic variables—in their operationalization of "good writing" in this writing portfolio assessment. This study provides a model for future validity inquiry of emerging context-rich writing assessment practices.

The best defense against inequitable assessment is openness. Openness about design, constructs, and scoring will bring out into the open the values, and biases of the test design process, offer and opportunity for debate about cultural and social influences, and open up the relationship between the assessor and the learner.

– C. Gipps

DOI: https://doi.org/10.37514/PER-B.2024.2166.2.02

An African American student came to the Writing Assessment Office at our western, land-grant public university and stated that she had heard that Black students failed our mid-career, university-wide Writing Portfolio at higher rates than other students. My office staff and I could not answer her because, since the program's inception in 1991, the Writing Assessment Office had never collected information regarding student race or ethnicity. The Writing Assessment Program fashioned itself as progressive: we administered a different kind of test than standardized ones so widely disparaged in writing circles. Our test was a portfolio that required students to turn in work produced for their regular coursework as well as complete an impromptu writing sample. A diagnostic evaluation was made by faculty from across the disciplines regarding the level of support needed for the student to successfully navigate the upper-division discipline-specific writing in the major courses required at our institution. Faculty raters used shared evaluation methodologies in which local context drives the articulation of assessment standards. As such, the connection between assessment, instruction, and curricular context was much stronger than standardized tests since much of the evaluation was based on coursework produced in undergraduate classroom settings, and the shared evaluation methodology relied on the expertise of classroom teachers in making these judgments. Students either passed the assessment or demonstrated a need for additional help, "Needs Work," mitigating the stakes for the test. The worst thing that happened to students was they were required to take structured instructional support as they navigated their upper-division writing requirements. The "Needs Work" rating did not follow the students: once they passed the additional coursework, the students' Writing Portfolio ratings were recorded as "Pass" on their university transcripts. In other words, students couldn't "fail" the Writing Portfolio. Program administrators tended to be satisfied with innovations developed for testing and contributions of the new shared evaluation rating procedures of our program, and adopted a stance consistent with other writing assessment scholars who claimed that "the advantages of portfolio assessment [had] overridden its problems, and as we [moved] into the twenty-first century portfolios achieved standing as the writing assessment method of choice" (White, 2005, p. 583). However, such a stance is detrimental to furthering an understanding of the complexity of shared evaluation practices in performance-based assessments and the effects they have on students. Schmidt and Camara (2004) confirmed the promise subscribed to performance assessments to

> reduce differences among groups because they provide students with hands-on opportunities to demonstrate their knowledge and understanding of how to solve problems

> rather than requiring students to simply recall facts. . . . Unfortunately, few large scale studies have examined differences among racial groups on performance assessments. (p. 193)

The one notable exception would be Breland et al.'s (2004) inquiry into the 'new' SAT which found "no significant prompt type effects for ethnic, gender or language groups, although there were significant differences in mean scores for ethnic and gender groups for all prompts" (p. 1). Cary-Lemon (2009) notes that "discourse about 'race' in [Composition Studies] reflects a fluctuating scholarly space" (W12), and argues for a self-critical look at the topics we have examined within our field related to race to see what has been included and excluded in our inquiries to examine these "reflections of racialized ideology over time" (W2).

While writing portfolio assessment tends to feel better to administrators and teachers, a limited number of quantitative or qualitative validational studies have been conducted through the revised framework of validity inquiry (AERA, APA, NCME, 1999; Kane, 2006). Such inquiries need to consider the interpretation and use of test scores as well as their consequences for students who take them. Kane (2006) asserts that validation "involves the development of evidence to support the proposed interpretations and uses [of test results] . . . to show that [such use] is justified . . . [and to assess] the extent to which the proposed interpretations and uses are plausible and appropriate" (p. 17). Perhaps owing to validity's psychometric roots, scholars in composition studies have had a general mistrust of validity research (Sharton, 1996; Lynne, 2004; Murphy, 2007). O'Neill (2003) documents how "validity has been—and continues to be—misconstrued in most of composition's assessment literature" (p. 49) highlighting the troubling "lack of rigorous composition research" (p. 51) into writing assessment methods regarding validity. Haswell (2005) also noted a general lack of replicable, aggregable, and data-driven scholarship in composition studies, characterizing the situation as an all-out war against this type of inquiry.

In spite of this, scholars have called for attentiveness to issues of validity in testing and assessment. Huot (1996) called for a "theory of writing assessment . . . [that recognizes] the importance of context, rhetoric, and other characteristics integral to a specific purpose and institution" (p. 552) and laid the groundwork for researchers to investigate composition-related issues of validity. The revised concept articulated in the Standards states that "validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests" (AERA, 1999, p. 9). Validity inquiries should include examinations of the consequences to the individuals taking the tests, and are no longer just comprised of different and individual components of validity (construct, content, predictive). The process of validation involves accumulating evidence to

provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself. When test scores are used or interpreted in more than one way, each intended interpretation must be validated. (AERA, p. 9) Kane (2006) asserted that

> validation employs two kinds of argument. An *interpretive argument* specifies the proposed interpretations and uses of test results by laying out the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances. The *validity argument* provides an evaluation of the interpretive argument. (p. 23)

The relevance of validity to writing assessment practitioners is apparent when validity is understood as an ongoing argument to be made rather than a static state to be achieved and justified. O'Neill (2003) contends that "validation arguments are rhetorical constructs that draw from all the available means of support" (p. 50). Huot and Schendel (1999) assert that validity and "assessment must be discussed in the context of ethics, for the consequences of assessment procedures are closely tied to the political and social contexts in which they take place" (p. 40). O'Neill (2003) argues that such lines of inquiry and research "[demonstrate] how systematic, ongoing validity research [function] to enhance a particular local test and contributes—both theoretically and practically to the scholarship of writing assessment" (p. 48). However, in spite of innovations and implementations of new contextually-based college writing assessment practices, systematic and rigorous validity inquiry into emerging college writing assessment practices have been limited.

O'Neill notes the reductive tendency in composition studies to simplify validity to mean "honesty . . . accuracy . . . and rightness" (2003, p. 49) that limits the complexity of the construct. There are many important theoretical calls for the discipline to wrestle with validity issues contextually or hermeneutically (Huot, 1996; Huot and Schendel, 1999; Moss, 1998a; Murphy, 2007; Inoue, 2007) and few forays of actual research and practice into validity inquiry in college writing assessment (Smith, 1993; Williamson and Huot, 1993; Haswell, 1998a and 2000; Broad, 2000; O'Neill, 2003; Hester, O'Neill, Neal, Edgington, & Huot, 2003; Elliot, Briller, & Joshi, 2007; Gere, Aull, Green and Porter, 2010). Researchers and scholars have neglected to conduct validity inquiries of locally developed writing assessment practices and so have not documented contributions or innovations these practices embody, and they fail to be attentive to students who take the exams. Kane (2006) says "there are, potentially, a large number of assumptions in any interpretive [validational] argument. We

take many of these assumptions for granted, at least until evidence to the contrary develops" (p. 23). To unearth some of these assumptions, previous scholars' criticism of standardized testing helps articulate where to begin: "what kind of proof do we have that students are wrong when they say, 'I don't belong in this dummy class'?" (Elbow, 1996, p. 93). While Elbow originally leveled this question at holistic or standardized tests, it is still relevant as a question for writing assessment programs that employ shared evaluation practices—locally developed, context-rich, practices that rely on faculty articulation values—whether via portfolios, direct-self placement, or other methods. Students who don't meet standards for writing tests face consequences that require completing additional coursework, spending additional time, spending additional money (perhaps), and dealing with the stigma of not passing the "test". Moss (1995) cites Cronbach and argues that "when the anticipated consequences [of assessment] 'impinge on the rights and life chances of individuals' (Cronbach, 1988, p. 6) . . . the investigation of consequences becomes particularly salient" (p. 11).

Rigorous validity inquiry allows for in-depth investigation of issues that we observe anecdotally—from student outrage at perceived unfair testing practices to patterns of course enrollment that may have more students of color populating the required writing support courses. Rigorous validity inquiry enables informed practice in a setting and directly addresses concerns of power highlighted by Huot and Williamson (1997) who note "assessment procedures [are] instruments of power and control, revealing so-called theoretical concerns as practical and political" (p. 44). They "fear that unless we make explicit the important power relationships in assessment, portfolios will fail to live up to their promise to create important connections between teaching, learning and assessing" (p. 44). Such a fear is applicable to any form of writing assessment that uses shared evaluation practices, particularly as these issues relate to test fairness. Camilli (2006) asserts while there are many aspects of fair assessment, it is generally agreed that tests should be thoughtfully developed and that the conditions of testing should be reasonable and equitable for all students . . . fairness issues are inevitably shaped by the particular social context in which they are embedded. (p. 221)

Certainly, as Schmidt and Camara (2004) observe, there have been "persistent score differences among racial groups" (p. 189) for a variety of standardized tests. Similar studies for performance-based assessments are still inconclusive but suggest that "subgroup gaps on traditional tests remain for [performance based] assessments" (p. 193). Most of this research, though, has occurred at the primary and secondary school level and not the college level.

Camilli (2006) states that "large differences are commonly encountered in test scores among groups of different races and ethnicities, and it is important to understand the extent to which these differences are artifacts of a test rather than

true proficiency" (p. 243). To address this, I conducted an empirically-based, descriptive validity inquiry into the large-scale writing portfolio assessment responsive to the African American student's question at my university. This inquiry begins by examining general performance trends by student race. It then conducts a sequential regression analysis into the construct of good writing as applied in the shared evaluation methodology used to assess the Writing Portfolio to identify the variables that raters actually use in the evaluation of students' writing, and to see if race is among them. This validity inquiry follows Moss' (2007) identification of

> productive directions for research in validity theory . . . [to develop] cases for validity research to both illustrate validity theory and to critique it . . . [including] cases as empirically based descriptions of the actual practices of working scientists, and . . . cases as critical analyses that locate our theories and practices in the sociohistorical-political contexts in which they are developed and used. (p. 96)

The question posed by the African American student regarding students of colors' performances on the Writing Portfolio opened up an avenue of research relevant for college writing assessment: Could the shared evaluation processes used by the university-wide Writing Portfolio assessment—and by other contextually defined writing assessment practices—be inadvertently complicit in perpetuating a system of discrimination? In other words, could teachers/evaluators unwittingly be disadvantaging students of color in a large-scale writing assessment program because of unstated biases related to race?

For this study, the operational definition of race is based upon the categories employed by my institution for collecting data related to race. These categories were based on an older definition of racial designations articulated by the federal Office of Management and Budget. These categories were not based on the most recent 1997 OMB revision to these designations articulated in Camilli (2006). The categories used in this study are American Indian or Alaska Native; Black or African American; Asian, Pacific Islander, Native Hawaiian; Hispanic or Latino; and White. This results in a less than nuanced view of race in this study, and, along with others, I recognize the limitations in such categorizations of race. Specifically, the American Anthropological Association (1997) asserted:

> Race and ethnicity both represent social or cultural constructs for categorizing people based on perceived differences in biology (physical appearance) and behavior. Although popular connotations of race tend to be associated with biology

> and those of ethnicity with culture, the two concepts are not clearly distinct from one another.

The APA Task Force on Diversity Issues at the Precollege and Undergraduate Levels of Education in Psychology (1998) argued that:

> "Race" has social meaning often accompanied by stereotyping; it suggests one's status within the social system and introduces power differences as people of different "races" interact with one another. 'Ethnicity,' on the other hand, connotes common culture and shared meaning. It includes feelings, thoughts, perceptions, expectations, and actions of a group resulting from shared historical experiences.

This study represents a starting point for this type of research, and hopefully future studies can include more complex representations of race and ethnicity.

VALIDITY INQUIRY AND WRITING PORTFOLIO INNOVATIONS

In the early 1990's, validation efforts for this program's Writing Portfolio focused on the Simple Pass methodology as this affected the largest number of students (about 60% of students who completed the Portfolio—roughly 2500 students out of 4200 who complete their Writing Portfolios each year), and at the time presented the most controversial and innovative contribution to the field of college writing assessment. The methodology of the writing assessment system represented a shift in writing assessment practices from holistic writing assessment—in which raters assigned an external numeric value to students' writing—to the expert-rater system (Haswell and Wyche-Smith, 1996; Haswell, 1998b; Haswell, 2001), a shared evaluation methodology which relies on context and teachers' judgments about students' abilities to manage the writing challenges of specific courses. The shared evaluation system used by our institution was based on the placement work of William Smith (1993) and was adapted to an upper-division context. Faculty raters review impromptu writing exams that sort writing obviously ready for upper-division writing intensive coursework from writing that was either very strong or very weak. Writing at either end of the spectrum—weak or strong—was sent on for further review and consultation by more experienced raters. The process assumed that additional focused rating time, information about the student's writing abilities through three additional course paper submissions, and faculty expertise would ensure the validity of the Writing Portfolio results. Virtually no validity attention was given to the results of "Needs Work," perhaps because such

widespread feeling existed among faculty about the poor quality of student writing. And, perhaps, the shift from holistically evaluating writing to relying on an innovative system of evaluation represented a significant enough move to not immediately surface new issues that embedded in the new methodology.

In response to the African American student's question, I investigated students of colors' performances on the Writing Portfolio for Academic Year 2004-05 according to the racial classifications collected by my institution. At that time, this institution reported the demographic profile of undergraduate students as 76 percent White; 1 percent American Indian Alaskan Native; 2 percent Black; 6 percent Asian Pacific Islander (API); 4 percent Hispanic; 3 percent non-resident aliens; and 8 percent unknown. Students' racial affiliation was obtained from this institution's Institutional Research Office by U.S. Census Bureau/ OMB categories, and then was combined with students' Writing Portfolio results. Tables 1.1 and 1.2 document the difference in performance percentages for the impromptu exam portion of the Writing Portfolio and the final review of the entire Writing Portfolio.

Tables 1.1 and 1.2 illustrate an unevenness in performance on the Writing Portfolio by race. Simply examining the percentages does not indicate whether these differences are significant. An analysis of variance was conducted on the performances of students on the timed writing portion of the Writing Portfolio by race. A random sample of 508 timed writing records were selected from the 5347 Writing Portfolio records recorded during AY 2004-2005. Students who spoke English as a second language were omitted from this analysis. An analysis of variance showed that the difference in performance by race on the timed writing portion of the Writing Portfolio was significant, $F(4, 503)=6.032$, $p=.000$. Post hoc analyses using Tukey's LSD for significance indicated that Black ($M=1.58$, $SD=.496$), API ($M=1.59$, $SD=.509$), and Hispanic ($M=1.7$, $SD=.462$) students' timed writing performances were significantly lower than White students ($M=1.84$, $SD=.550$).

Table 1.1. Comparison of Performance Rates on the Writing Portfolio Impromptu Exam by Race

Population	Pass	Distinction	Needs Work
Combined—all students	58.8%	8.6%	32.6%
American Indian	52.2%	8.7%	39.1%
API	47.3%	6.1%	46.6%
Black	48.6%	4.3%	47.1%
Hispanic	55.7%	5.1%	39.2%
White	60.3%	8.2%	31.5%

Note. Source: Writing Assessment Office, Database, (AY 2004-2005)

Table 1.2. Comparison of Performance Rates on the Final Writing Portfolio Review by Race

Population	Pass	Distinction	Needs Work
Combined—all students	78.1%	8.6%	13.3%
American Indian	82.6%	8.7%	8.7%
API	71.6%	5.3%	22.9%
Black	77.6%	2.9%	20.0%
Hispanic	82.3%	3.8%	13.9%
White	82.4%	6.9%	10.7%

Note. Source: Writing Assessment Office, Database, (AY 2004-2005)

Additionally, a second ANOVA was run to compare students' performances by race for the final Writing Portfolio review. A random sample of 749 final Writing Portfolio performances by race was selected from the 5378 available records for AY 2004-2005. Again, multi-lingual speakers were omitted from this analysis. The results indicated a significant difference in the performance on the final Writing Portfolio by race, $F(4, 744)=3.120$, $p=.015$. Post hoc analyses using Tukey's LSD for significance indicated that Black students ($M=-1.81$, $SD=.429$) performed significantly lower on the final Writing Portfolio review than all other students: American Indian ($M=2.00$, $SD=.434$), API ($M=1.97$, $SD=.412$), Hispanic ($M=1.93$, $SD=.411$), and White ($M=1.94$, $SD=.493$).

An analysis that ended here would purely speculate about the reasons underlying the differences in performance by race and wouldn't address "the extent to which these differences are artifacts of a test rather than true proficiency" (Camilli, p. 243) or whether they are result of a "construct-irrelevant variance [which] refers to the degree to which test scores are affected by processes that are extraneous to its intended construct" (AERA et al., 1999, p. 10). The *Standards* (1999) state:

> The idea that fairness requires overall passing rates to be comparable across groups is not generally accepted in the professional literature. Most testing professionals would probably agree that while group differences in testing outcomes should in many cases trigger heightened security for possible sources of test bias, outcome differences across groups do not in themselves indicate that a testing application is biased or unfair. (AERA, 1999, p. 75)

Breland et al.'s (2004) study approached testing difference by race from the perspective of reliability, but as Broad (2000) noted, writing assessment scholars

tend to feel a tension between what he characterized as 'validity and reliability debates' that occur between positivistic and hermeneutic traditions. Complicating this issue further, the Writing Portfolio uses non-parametric data for its system of measurement—ratings are recorded as Needs Work, Pass, or Pass with Distinction—and so have limited transferable numeric value resulting in equally limited statistical analyses. The question raised by the African American student was apt because it highlighted our own program's general tendency—as well as that of composition studies—to neglect to attend to students of colors' experiences in our writing assessment systems.

In composition studies, there has been a great deal of agenda setting and calls for research regarding potential biases against students of color in writing assessment practices, (Farr and Nardini, 1996; Lippi-Green, 1997; Mountford, 1999; Hamp-Lyons and Condon, 2000; Murphy, 2007) but no empirical or qualitative inquiry into students of color's actual experiences in college-level writing assessment systems. Farr and Nardini (1996) suggest that a dominant paradigm of writing instruction exists called "essayist literacy" in which "high value is placed on language, either oral or written, that is rational, decontextualized, explicit, and carefully ordered internally" (p. 108), "[and] . . . the social and cultural mindset that construes rationality, explicitness and order as fundamental values of literate text—namely, the (primarily white and male) Anglo-American analytic orientation" (p. 117).

Other researchers note possible deleterious effects of race applicable to context-rich assessment situations. Omi and Winant (1994) describe how racial formation occurs in everyday face-to-face experience in "the many ways in which, often unconsciously, we 'notice' race . . . One of the first things we notice about people when we meet them (along with their sex) is their race. We utilize clues about *who* a person is" (p. 59). They argue that "our ability to interpret racial meanings depends on preconceived notions of a racialized social structure. . . . We expect people to act out their apparent racial identities" (p. 59). In an assessment system predicated on faculty articulation of values, could Writing Portfolio raters have unstated expectations for student writing and who they think might "write" like students of color resulting in biased assessment of their writing? Moss and Shutz (2001) assert "even in the most intimate settings, issues of inequality, cultural and racial difference, gender, and class affect dialogues in subtle ways giving some voices more authority while silencing others" (p. 42). Ball (1997) concluded that holistic writing assessment procedures used in middle schools disadvantaged African American students because they did not share the same linguistic features as middle-class, Anglo American students and the middle class European teacher who evaluated their writing. In an assessment context, such findings are troublesome because these instances suggest an unfair educational

system and that the assessments may be perpetuating distressing consequences on particular groups of students who take these tests. Bond (1995) argued that

> performance assessments are, at least potentially, less biased and more fair to traditionally disadvantaged students because such tests, when properly used, can merge instruction and assessment rather than test abilities . . . that are only remotely connected to the everyday experience of these students (p. 21).

Bond concurs with writing assessment researchers who tout the value of portfolio assessment, but warns that performance assessments still have significant unresolved issues regarding bias and validity. The lack of straightforward validity evidence for portfolio assessment is corroborated by LeMahieu et al. (1995) and Griffee (2002). In particular, Bond cautions that examination of consequential aspects of validity should "not only [include] the elimination of elements in assessment that unduly *disadvantage minority persons* but also the elimination of construct-irrelevant elements that may subtly *advantage majority persons over others"* (p. 23) by asserting:

> People also hold purely prejudicial beliefs that can affect their objective assessment of others' ability . . . it would take an extraordinary effort on my part to give the same evaluation to two individuals who are identical in every way except that one has a high British accent, and the other a deep southern drawl! (pp. 23-24)

Taken in the context of a shared evaluation setting, Bond implies the potential for raters to privilege or diminish students' writing based on how the writing fits a pre-conceived notion of 'good writing' and that this definition of good writing may be susceptible to bias.

Moss (1998a) critiqued limitations in our program's early forays into validity inquiry of the junior-level Writing Portfolio assessment (Haswell 1998a) advocating that our program consider "to what extent . . . the writing program [is] complicit in simply reproducing a narrow model of academic writing (and the understanding of knowledge it entails) without providing opportunity for the values implicit in the model to be illuminated and self-consciously considered" (p. 120). Moss (1998b) argues and Huot and Schendel (1999) later reiterate that "we need to study the actual discourse and actions that occur around products and practices of testing" (p. 7). In a shared evaluation system, then, a primary validity focus should be on how faculty articulate and operationalize the standards of good writing for the particular context. For this study, students' readiness for upper-division, disciplinary-specific writing in the major work represents that

context. The Writing Portfolio assessment requires the articulation of commonalities of student readiness requisite for their entry into diverse, disciplinary-specific discourse communities. Given the differences in the general Writing Portfolio performance data between different racial groups, consideration of race is key in how faculty operationalize standards for good writing.

METHODS

This validity inquiry examines how raters functionally define good writing through sequential regression analysis techniques that examine actual student products submitted for the Writing Portfolio. These analyses are conducted through three frameworks: the Writing Portfolio assessment criteria, an Alternate set of writing criteria, and Demographic criteria. Each of these frameworks are applied to the two distinct writing tasks—impromptu writing and coursework written for regular undergraduate courses across the disciplines—selected for inclusion in the Writing Portfolio as representative of the student's best writing.

This inquiry allows for more sophisticated statistical analysis of the factors that account for the variability in the writing quality scores of the Writing Portfolio using a finer grained instrument, the Writing Portfolio Differential Scale for Writing and Demographic Information (see Appendix A). The Writing Portfolio Differential Scale was developed by this researcher to interpret raters' evaluation behaviors and determine the criteria they seemed to actually use to evaluate writing; the criteria that seemed to carry more weight in their evaluation process; and whether demographic features perceived about writers accounted for any part of the evaluation results. Guiding questions for this inquiry include:

1. What is the definition of "good writing" that faculty raters apply when evaluating the Writing Portfolio?
2. Do faculty raters make demographic assumptions about students based on their writing that effect the results?
3. Does this evaluation privilege forms of writing according to race?

Sequential regression analysis is used to assess the relationship between a dependent variable (like writing quality) and several independent variables (like criteria that comprise quality—focus, organization, or use of Standard American English and so on) by entering variables in a specific order into regression equations to identify which variables account for the variability in—or the criteria that comprise—the overall score (Tabachnick & Fidell, 2006). The methodology for this study was piloted in an earlier project by the researcher in which

the Writing Portfolio Differential Scale was tested and the order of the criteria variables were established for the regression analysis (Kelly-Riley, 2006).

The Writing Portfolio Differential Scale for Writing and Demographic Information was adapted from the work of Piche, Rubin, Turner, and Michlin (1978) and Osgood (1957). Piche et al. used the work of Osgood to examine whether teachers evaluated Black elementary students' writing differently from their White counterparts. Osgood created semantic differential scales that "relate to the functioning of representational processes in language behavior and hence may serve as an index of these processes" (p. 9). Osgood's work developed out of experimental psychology to establish pairs that exist in what he called semantic space, "which are assumed to represent a straight line function that passes through the origin of this space, and a sample of such scale then represents a multidimension space" (p. 25). His work attempts to quantify the complexity inherent in measuring a construct like writing. Piche et al. (1978) developed their scale items based on the research of Osgood (1957) and the application of these scales by Williams, Whitehead, and Miller (1971) who examined relationships of attitudes and children's speech. Piche et al. examined teachers' responses to their scale items by presenting teachers with different samples of writing. Some of the samples contained inserted types of speech the researchers characterized as African American Vernacular English (AAVE). Actual samples of Black students' writing were not used for their study. Instead, they used a piece of writing, and added features identified as AAVE into the text.

In addition, Rubin and Williams-James (1997) examined the ways teachers responded to international students' writing using similar scales. These researchers created a text and inserted types of speech that appeared to be consistent with writers from different nationalities. They did not use actual student products for their evaluation. The instrumentations of these differential scales, however, set a precedent to examine instructors' impressions of students' writing.

For this study, the Writing Portfolio Differential Scale contains three separate criteria frameworks: Writing Portfolio Criteria, the programmatic areas articulated, published and evaluated by the Writing Program (Comprehension of the Task, Focus, Organization, Support, and Proofreading) and two other frameworks of criteria—Alternate Writing Criteria and Demographic Criteria—which were previously used by Piche et al. and Rubin and Williams-James. The Alternate Writing Criteria include Coherence, Use of Standard American English, Logic, Grammar, Creativity, Level of Language Passivity, and Quality of Writing; and the Demographic Criteria include the rater's perception of the writer in many areas: Strength of Writer, Intelligence, Socio-Economic Status, Level of Cultural Advantage, Confidence, and Comfort as a Writer. This study examined actual samples of student writing composed by actual students for

undergraduate courses across the disciplines subsequently submitted for their university required Writing Portfolios.

A group of faculty Writing Portfolio raters were trained to apply the different variables of the Writing Portfolio Differential Scale to individual components of students' Writing Portfolio submissions. Two rating sessions were held in the fall of 2006, and consisted of thirty-three raters. Four raters (12%) were tenure-line faculty; thirteen (39%) were adjunct faculty; fourteen (42%) were graduate teaching assistants with extensive teaching experience; and the other two raters (6%) were other position classifications. Of this group, 24 were white; 4 multiracial; 1 Hispanic; 3 Asian Pacific Islander; and 1 African American. Eighteen percent were multi-lingual and 82% were native speakers of English. Seventy-six percent of the raters were female and 24% of the raters were male.

Two hundred and fifty writing portfolios were selected for this study—fifty from each racial classification used by the researcher's institution (American Indian Alaska Native, Asian Pacific Islander, Hispanic, Black, and White). The selected Writing Portfolios had been submitted between 2003-2006. Each Writing Portfolio contained an impromptu exam and three course papers. Overall, one thousand samples of writing were evaluated for this study.

The analysis examined a random sample of 150 impromptu exams and 300 individual course paper submissions. The samples were selected by using the SPSS option to create a randomized list for data analysis. In order to instill confidence in the results, sample sizes for the analyses followed Tabachnick and Fidell's (2005) "simple rules of thumb [for sample size for regression analyses]: N>50+8m (where *m* is the number of [independent variables] for testing the multiple correlation" (p. 123). Shavelson's (1996) rule of thumb encouraged at least fifty subjects, and ten times as many cases as independent variables, which would be at least 120 for each analysis. Again, separate regression analyses were conducted for the three different frames for the impromptu exams and for the three different frames for the course paper submissions. Each analysis conducted a sequential regression analysis on each of the scale frameworks: Writing Portfolio criteria, Alternate Writing criteria, and Demographic criteria. In other words, a total of six regression equations were calculated: three sequential regression equations were established for each criteria framework for each type of writing resulting in equations that account for the variability of writing quality scores.

Sequential regression analyses were conducted on a random sample of 150 Writing Portfolio impromptu exams to account for the variability in the quality of the writing through the Writing Portfolio criteria, Alternate Writing criteria, and Demographic criteria. Each variable was rated on a scale from 1 to 6.

Students' racial identities were converted to a nominal scale (American Indian=1; African American=2; Asian Pacific Islander=3; Hispanic=4; White=5) and were entered first into each regression equation. A sequential regression analysis was conducted and the entry order of the variables was based upon the stepwise regression analysis results from Kelly-Riley (2006). Table 1.3 details the order of the criteria variables were entered into the sequential equation as well Cronbach's alpha.

Table 1.3. Variable Entry Order into the Sequential Regression Analysis and Reliability Data for the Timed Writing analysis

Criteria Framework	Variable Order of Entry	Cronbach's Alpha
Writing Portfolio	Race, Focus, Proofreading, Support, Comprehension of task, and Organization	.8946
Alternate Writing	Race, Coherence, Logic, Creativity, Grammar, use of Standard American English, Language passivity	.8902
Demographic	Race, and raters' perceptions of writers' Confidence, Intelligence, Comfort with writing, Socio-economic status, and Cultural advantage	.7902

Note: N=150

Table 1.4 details the order of the criteria variables were entered into the sequential equation as well Cronbach's alpha. The entry order of the variables are slightly different based on the results from the stepwise analysis conducted by Kelly-Riley (2006).

Table 1.4. Variable Entry Order into the Sequential Regression Analysis and Reliability Data for the Course Paper analysis

Criteria Framework	Order of Entry	Cronbach's Alpha
Writing Portfolio	Race, Focus, Proofreading, Support, Organization and Comprehension of task	.8512
Alternate Writing	Race, Coherence, Logic, Grammar, Creativity, Use of Standard American English, Language passivity	.8647
Demographic	Race, and raters' perceptions of writers' Comfort with writing, Intelligence, Confidence, Socio-economic status, and Cultural advantage.	.8560

Note: N=300

RESULTS

The results from the sequential regression analyses for both types of writing—impromptu and course paper submissions—suggest that the definition of good writing is based more on variables of coherence, correctness, and confidence as applied by faculty raters in this large scale Writing Portfolio assessment. Race did not contribute significantly to faculty raters' functional definition of "good writing" for any of the frameworks whether in the timed exam format or for the course papers. Surprisingly, faculty raters operationalize their assessment of "good writing" based on criteria accounted more through non-programmatic evaluation criteria of the Alternate Writing framework variables. In addition, a high percentage of the writing scores—for impromptu writing as well as papers written for courses—included demographic considerations of the writer. Higher percentages of writing quality scores were accounted for through coherence and grammar, part of the Alternate Writing framework. These variables overlap with focus and mechanics, which account for writing quality through the Writing Portfolio framework, although the Writing Portfolio variables account for a slightly lesser percentage of the writing quality score. A surprisingly high percentage—nearly two thirds of the score—of writing quality is also accounted for through raters' perceptions of student writers' intelligence and comfort with writing. For each type of writing—impromptu exams and course paper submissions—race did not contribute significantly to writing quality score through any of the frameworks.

Table 1.5 details the separate regression equations that account for the variability in the impromptu exam analysis. The Alternate Writing Criteria accounted for the most variability in the impromptu writing score.

Findings for the Impromptu Exam Analysis

Table 1.5. Raw and Standardized Regression Coefficients and Percent of Variance in Timed Writing Quality explained by (A) Writing Portfolio Criteria, (B) Alternate Writing Criteria and (C) Demographic Criteria

Significant Regression Equations	B	b	% of variance explained
(A) Focus + Mechanics +-Support	.198	.224**	61.6
(B) Coherence+ Creativity + Grammar	.223	.236**	68.0
(C) Intelligence+ Comfort	.546	.506**	60.9

*Note. Each frame represents a separate regression equation with the variables in the order in which the regression analysis specified. N=150 *p<.05. **p<.01.*

Tables 1.6, 1.7, and 1.8 provide detailed analysis of the significant regression equations and the differences in the percentages of the variance explained for each of the three separate criteria frameworks.

Table 1.6. Raw and Standardized Regression Coefficients and Percent of Variance in Timed Writing Quality explained by Writing Portfolio Criteria

Predictor Variables/Regression Equations	B	b	% of Variance Explained
Focus	.214	.235**	38.1
Focus + Mechanics	.467	.492**	59.5
Focus + Mechanics+ Support	.198	.224**	61.6

*Note. N=150 *p<.05. **p<.01.*

Table 1.7. Raw and Standardized Regression Coefficients and Percent of Variance in Timed Writing Quality explained by Alternate Writing Criteria

Predictor Variables/Regression Equations	B	b	% of Variance Explained
Coherence	.518	.512**	62.0
Coherence +Creativity	.218	.208**	65.5
Coherence +Creativity +Grammar	.223	.236**	68.0

*Note. N=150 *p<.05. **p<.01.*

Table 1.8. Raw and Standardized Regression Coefficients and Percent of Variance in Timed Writing Quality explained by Demographic Criteria Variables

Predictor Variables/Regression Equations	B	b	% of Variance Explained
Intelligence	.533	.396**	42.0
Intelligence +Comfort	.546	.506**	60.9

*Note. N=150 *p<.05. **p<.01.*

Findings for the Course Paper Analyses

More of the variance in the writing quality score was accounted for in the assessment of the course papers. Table 1.9 details the separate regression equations that account for the variability in the course papers. The Alternate Writing criteria accounted for the most variability in the course paper review.

Table 1.9. Raw and Standardized Regression Coefficients and Percent of Variance in Course Paper Writing Quality explained by (D) Writing Portfolio Criteria, (E) Alternate Writing Criteria and (F) Demographic Criteria

Significant Regression Equations	B	b	% of Variance Explained
(D) Focus + Mechanics + Organization	.198	.188**	72.0
(E) Coherence + Logic + Grammar	.420	.452**	77.0
(F) Comfort + Intelligence + Confidence	.187	.158**	64.1

*Note. Each frame represents a separate regression equation with the variables in the order in which the regression analysis specified. N=300 *p<.05. **p<.01.*

Tables 1.10, 1.11 and 1.12 provide detailed analysis of the significant regression equations and the differences in the percentages of the variance explained for each of the three separate criteria frameworks for the review of the course papers.

Table 1.10. Raw and Standardized Regression Coefficients and Percent of Variance in Course Paper Writing Quality explained by Writing Portfolio Criteria

Predictor Variables/Regression Equations	B	b	% of Variance Explained
Focus	.253	.222**	38.2
Focus + Mechanics	.545	.592**	70.8
Focus + Mechanics Organization	.198	.188**	72.0

*Note. N=300 *p<.05. **p<.01.*

Table 1.11. Raw and Standardized Regression Coefficients and Percent of Variance in Course Paper Writing Quality explained by Alternate Writing Criteria

Predictor Variables/Regression Equations	B	b	% of Variance Explained
Coherence	.421	.360**	64.2
Coherence+ Logic	.192	.162**	68.2
Coherence+ Logic+ Grammar	.420	.452**	77.0

*Note. N=300 *p<.05. **p<.01.*

Table 1.12. Raw and Standardized Regression Coefficients and Percent of Variance in Course Paper Writing Quality explained by Demographic Criteria Variables

Predictor Variables/Regression Equations	B	b	% of Variance Explained
Comfort	.465	.426**	55.6
Comfort + Intelligence	.369	.318**	63.5
Comfort + Intelligence +Confidence	.187	.158**	64.1

*Note. N=300 *p<.05. **p<.01.*

DISCUSSION AND IMPLICATIONS

The first research question focused on the definition of good writing used by faculty raters and the results from the six separate regression analyses show the surprising ways that faculty operationalize this construct. First, a comparison of the two Writing frameworks (Writing Portfolio criteria and Alternate Writing criteria) show that coherence, focus, and correctness all contribute significantly to the functional definition of "good writing" applied by faculty raters in the context of this mid-career diagnostic assessment. All of these variables contribute significantly to writing quality in both the impromptu writing situation and for the course paper evaluation (in which students theoretically would have time to plan, draft, and revise). More variance in writing quality is accounted for in the evaluation of the course papers than the impromptu exams. Nearly a third of the impromptu writing quality score is unaccounted for while less than a quarter is unaccounted for in the course papers.

Secondly, raters seem to apply more non-programmatic variables not overtly articulated by the Writing Program. More of the variance in the writing quality score—for both impromptu writing and course papers—is accounted for by the non-programmatic Alternate Writing criteria. The variable of Coherence accounts for the largest percentage of the variance in which Coherence, by itself, accounts for 62% of timed writing quality and 64.2% of course paper writing quality. On the other hand, Focus, as a standalone variable, accounts for only 38.1% of the variance of timed writing quality and 38.2% of course paper quality. For impromptu writing, Creativity is a variable considered by raters whereas the Writing Portfolio criteria include Support. These two variables are dissimilar. However, there is some overlap between the variables of the two frameworks of writing criteria as Mechanics and Grammar are included in all four regression equations, and logic and organization are similar variables included in the course papers frameworks. In spite of the published and articulated Writing Portfolio criteria, raters

seem to apply idiosyncratic criteria that fall outside of the intended assessment. Perhaps this disconnect can be explained by the explicit instructions in the rating sessions for raters to reference their classroom writing experiences and expectations and to be guided by the Writing Portfolio criteria in the assessment. They are asked to operationalize the criteria as relevant to their disciplinary realities.

Similarly surprising, raters evaluate impromptu writing with slightly different expectations than the writing done within courses. While Focus and Mechanics are included both Writing Portfolio criteria, Support is used by raters to assess impromptu writing quality while Organization replaces it in the evaluation of the course papers. Likewise, this trend is observed in the Alternate Writing criteria. Coherence and Grammar account for the variance in writing quality for impromptu writing and course papers, but Creativity is important in impromptu writing whereas Logic replaces it in the course paper writing. These results suggest that faculty have different expectations for the two different writing tasks included in the same Writing Portfolio. These results do not differentiate between one set of criteria being better than the other; they only indicate that faculty seem to view these tasks differently. Certainly, this interesting result deserves further study.

The second research question examines whether the operationalized definition of good writing included demographic information. The findings suggest that large percentages of the variance of writing quality are accounted for through the Demographic framework—primarily through the rater's perception of the writer's intelligence and comfort with writing. The variables of race, perceived economic status, and perceived cultural advantage did not contribute significantly to the writing quality score. While the two writing frameworks have more obvious overlap, the demographic criteria seem to overlap with writing issues too. The demographic criteria that faculty use to account for writing quality are based on variables that would be reasonable to identify a writer as needing help: the student's comfort level with writing, the student's confidence with writing, and the teacher's perception of the student's intelligence. The variables are not related to demographic features that are irrelevant to the classroom.

The third question examined whether the assessment process privileged forms of writing according to race. The findings from this study suggest that race is not a significant contributor to the faculty's assessment of students' writing for either the impromptu writing or papers written for courses. The results of the sequential regression analyses suggest that race does not significantly account for the variance in good writing. However, students' performances by race on the Writing Portfolio are significantly different like the studies conducted by Schmidt and Camara (2004) and Breland et al. (2004), but the rating processes used by faculty raters do not seem to the cause for these differences.

Such concern about the relationship between the rater and the writer is warranted. Ball (1997) documented potential bias by readers for writers based on dissimilar cultural backgrounds. Smitherman's extensive research (highlighted in Smitherman and Villanueva, 2003) has documented different linguistic structures of African American students and their implications in educational settings. This study, though, found somewhat different results. The rating corps used for this study represented a linguistically and culturally diverse set of faculty—who were also representative of the regular Writing Portfolio rating corps—attempting to address some of the concerns raised by Ball. Admittedly, this study did not intend to address the specific relationship between rater and writer.

Furthermore, the extent to which Mechanics contributes to writing quality is interesting in the light of Smitherman's research, but this study included more racial categories than Smitherman's studies, which focused primarily on African Americans. Such distinct differences between raters and writers with a multitude of different backgrounds may not be as detectable as comparisons that look at only two racial groups. Even though race was not a variable that accounted for any writing quality in this study, some of Smitherman's findings that connect race and linguistic structure might seem supported by this study. Specifically, Mechanics accounts for a great deal of the Writing Portfolio quality score. Mechanics accounts for 32.6% to the variance in impromptu writing and 21.4% of the course papers. Overall, though, the Writing Portfolio criteria account for less of the writing quality than the Alternate Writing criteria. In the Alternate criteria, Grammar, while a significant contributor to quality, did not account for as much as Mechanics, with 8.8% of the variance explained for impromptu writing and 2.5% in the course papers. Issues of Coherence that go beyond Grammar seem to be more important in raters' assessment of students' writing.

While the findings from this study suggest that race does not contribute significantly to raters' operationalization of good writing, it is disconcerting that there are statistically significant differences in performances by race on the Writing Portfolio. While the reason may not be in how the raters evaluate student writing, the subject requires further investigation. Schmidt and Camara (2004) summarize the prevailing theories used to explain the gap in standardized test performances by race: inequitable educational preparation, poverty, discrimination, poor educational opportunities, and lack of access to educational resources. Studies such as these would be useful in large-scale performance-based assessment programs, and studies examining the effectiveness of the structured support required by these programs would be the next logical step in validity research.

Validity, again, refers to the use and interpretation of test scores in a particular setting. What do these results mean for the use and interpretation of the test

scores in the university-wide Portfolio context? The purpose of the Writing Portfolio is to assess students' readiness for the upper-division Writing in the Major courses. In the rating sessions, faculty are overtly asked to draw on their classroom experiences and expectations for the assessment situation. Perhaps this request for raters to draw on pedagogical reference points helps explain the large role that the Alternate Writing criteria play in accounting for writing quality. These findings are consistent with Broad's qualitative study that document the frustration faculty felt in rubric-based assessments in first year writing programs. Since the Writing Portfolio relies on the multitude of disciplinary definitions of good writing, it's important to have the starting point of common language articulated in the Writing Portfolio criteria and to begin to fully acknowledge the additional role that other non-programmatic criteria play. Additionally, in this Writing Portfolio system, frustration levels are mitigated in that faculty don't have to agree about a static definition of writing; faculty simply place student writers into three broad categories of placement: Pass, Pass with Distinction, and Needs Work. The broadly defined functional placements mask the complex process behind the rating behaviors. These rating behaviors need to be routinely examined.

These findings suggest that writing assessment program administrators need to play more of active role in looking at published program criteria, standards for writing, and faculty enactment of these standards. The focused time allotted to the evaluation of writing at the ends of the spectrum (weak or strong) in the shared evaluation, expert-rater system does not translate into a systematic application of the criteria of good writing as articulated through the programmatic rubric. Locally developed writing assessment programs—whether portfolios or directed self-placement or other mechanisms which rely on faculty articulation of standards—need to compare the published criteria used by their programs to the criteria used functionally through the rating process. This point is the "significant site of power and knowledge" (O'Neill, 2003, p. 62) so often ignored by compositionists.

A tension exists between the criteria the Writing Program articulates and publishes, and the actual multi-dimensional criteria enacted by faculty raters. The ways in which programmatic criteria and disciplinary expectations intersect must be examined further because they most certainly inform and reform each other in a system that intends to be responsive to validity and reliability concerns. The absence or limited contribution of the some of the programmatic Writing Portfolio criteria—comprehension of the task, organization, and support—to the writing quality score points to a disjuncture. The findings suggest that these criteria—as faculty use them—either contribute minimally to the quality score or not at all for both the impromptu and course paper evaluations. This omission raises the

question as to whether raters—who are hired for these positions based on their extensive teaching expertise—don't value or don't know how to evaluate for these criteria areas. While the program advertises and publishes specific criteria for evaluation of Writing Portfolios, more than half of these criteria areas are not utilized by raters in the evaluation setting. This omission questions the extent to which these criteria are employed in classroom settings. The program administrators must be aware of the tendency of raters to draw on personal pedagogical expectation, and to move the raters toward the programmatic criteria particularly for decisions that fall on the ends of the spectrum. Improved rater training and overt conversations about this tendency in norming sessions might be a way to begin to further identify and address these issues.

Finally, research that includes more nuanced considerations of race and ethnicity into these large scale writing assessment practices need to be more commonplace. Educational research already has a robust agenda of research in standardized tests related to race and ethnicity, but most times, the standardized tests are separated from the instructional or local context. Composition studies needs to embrace a similar research agenda which considers the hermeneutically-oriented assessment approaches that are rooted in local context. While this study only examines the construct of good writing as applied by raters, there are many other angles of necessary research and validity inquiry for students of colors' experiences in context-rich, locally-developed writing assessment programs. Given the more mainstream position that college writing assessment methodologies have garnered of late, such inquiry is important, timely, and vital—not only to examine the quality of the practices, but to ensure that such methodologies are not intentionally or unintentionally leveling consequences for students—particularly those represented by small populations who may be easily overlooked.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. American Educational Research Association.

American Anthropological Association (1997). Response to OMB directive 15. *Race and ethnic standards for federal statistics and administrative reporting*.

American Psychological Association Task Force on Diversity Issues at the Precollege and Undergraduate Levels of Education in Psychology (1998). Enriching the focus on ethnicity and race. *Monitor, 29*(3).

Ball, A. (1997). Expanding the dialogue on culture as a critical component when assessing writing. *Assessing Writing, 4*(2), 169-202.

Bond, L. (1995). Unintended consequences of performance assessment: Issues of bias and fairness. *Educational Measurement: Issues and Practice*, *14(*4), 21-24.

Breland, H., Kubota, M., Nickerson, K., Trapani, C., Walker, M. (2004). *New SAT writing prompt study: Analyses of group impact and reliability* (Report No. 2004-1). College Entrance Examination Board.

Broad, B. (2000). Pulling your hair out: Crises of standardization in communal writing assessment. *Research in the Teaching of English, 35*(2), 213-260.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed) (pp. 221-256). American Council on Education/Oryx Press Series on Higher Education.

Clary-Lemon, J. (2009). The racialization of composition studies: Scholarly rhetoric of race since 1990. *College Composition and Communication, 61*(2), W1-W17.

Cronbach, L. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3-17). Erlbaum.

Elbow, P. (1996). Writing assessment in the 21st century: A utopian view. In L. Z. Bloom, D. A. Daiker, & E. M. White (Eds.), *Composition in the twenty-first century: Crisis and change* (pp. 83-100). Southern Illinois University Press.

Elliot, N., Briller, V., & Joshi, K. (2007). Portfolio assessment: Quantification and community. *Journal of Writing Assessment, 3*(1), 5-30. https://escholarship.org/uc/item/8nm1m6xc

Farr, M. & Nardini, G. (1996). Essayist literacy and sociolinguistic difference. In E. M. White, W. D. Lutz & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 108-119). Modern Language Association.

Gere, A. R., Aull, L., Green, T., and Porter, A. (2010). Assessing the validity of directed self-placement at a large university. *Assessing Writing, 15(3)*, 154-176.

Griffee, D. (2002). Portfolio assessment: Increasing reliability and validity. *The Learning Assistance Review: The Journal of the Midwest College Learning Center Association, 7*(2), 5-17.

Hamp-Lyons, L. and W. Condon. (2000). *Assessing the portfolio: Principles for practice, theory and research*. Hampton Press.

Haswell, R. (1998a). Multiple inquiry in the validation of writing tests. *Assessing Writing*, *5*(1), 89-109.

Haswell, R. (1998b). Rubrics, prototypes and exemplars: Categorization and systems of writing placement. *Assessing Writing*, *5*(2), 231-268.

Haswell, R. (2000). Documenting improvement in college writing: A longitudinal approach. *Written Communication*, *17(3)*, 220-236.

Haswell, R. (Ed.). (2001). *Beyond outcomes: Assessment and instruction within a university writing program*. Ablex.

Haswell, R. (2005). NCTE/CCCC's recent war on scholarship. *Written Communication*, *22*(2), 198-223.

Haswell, R, & S. Wyche. (1996). A two-tiered rating procedure for placement essays. In T. W. Banta (Ed.), *Assessment in practice: Putting principles to work on college campuses* (pp. 204-207). Jossey-Bass.

Hester, V., O'Neill, P., Neal, M., Edgington, A., & Huot, B. (2007). Adding portfolios to the placement process. In P. O'Neill (Ed.), *Blurring boundaries: Developing writers, researchers, and teachers* (pp. 61-90). Hampton Press.

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

Huot, B. & Schendel, E. (1999). Reflecting on assessment: Validity inquiry as ethical inquiry. *Journal of Teaching Writing*, *17*(1-2), 37-55.

Huot, B. & Williamson, M. M. (1997). Rethinking portfolios for evaluating writing: Issues of assessment and power. In K. B. Yancey and I. Weiser (Eds.), *Situating portfolios: Four perspectives* (pp. 43-56). Utah State University Press.

Inoue, A. (2007). Articulating Sophistic rhetoric as a validity heuristic for writing assessment. *Journal of Writing Assessment*, *3*(1), 31-54. https://escholarship.org/uc/item/64n8z5mz

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.) *Educational measurement* (4th ed.) (pp. 17-64). American Council on Education/Oryx Press Series on Higher Education.

Kelly-Riley, D. (2006). A validity inquiry into minority students' performances in a large-scale writing portfolio assessment. (Doctoral Dissertation, Washington State University).

LeMahieu, P. G., Gitomer, D. H., & Eresh, J. T. (1995). Portfolios in large-scale assessment: Difficult but not impossible. *Educational Measurement: Issues and practice*, *14*(3), 11-28. https://doi.org/10.1111/j.1745-3992.1995.tb00863.x

Lippi-Green, R. (1997). *English with an accent: Language, ideology, and discrimination in the United States*. Routledge.

Lynne, P. (2004). *Coming to terms: A theory of writing assessment*. Utah University Press.

Moss, P. A. (1995). Themes and variations in validity theory. *Educational Measurement: Issues and Practice*, *14*(2), 5-13.

Moss, P. A. (1998a). Testing the test of the test: A response to "Multiple inquiry in the validation of writing tests." *Assessing Writing*, *5*(1), 111-122.

Moss, P. A. (1998b). The role of consequences in validity theory. *Educational Measurement: Issues and Practice*, *17*(2), 6-12. https://doi.org/10.1111/j.1745-3992.1998.tb00826.x

Moss, P. A. (2007). Joining the dialogue on validity theory in educational research. In P. O'Neill (Ed.), *Blurring boundaries: Developing writers, researchers, and teachers* (pp. 91-100). Hampton Press.

Moss, P. A & Schutz, A. (2001). Educational standards, assessment and the search for consensus. *American Educational Research Journal*, *38*(1), 37-70.

Mountford, R. (1999). Let them experiment: Accommodating diverse discourse practices in large-scale writing assessment. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: The role of teachers' knowledge about text, learning, and culture* (pp. 366-396). National Council of Teachers of English.

Murphy, S. (2007). Culture and consequences: The canaries in the coal mine. *Research in the Teaching of English*, *42*(2), 228-244.

Omi, M. & Winant, H. (1994). *Racial formation in the United States: From the 1960's to the 1990's*. Routledge.

O'Neill, P. (2003). Moving beyond holistic scoring through validity inquiry. *Journal of Writing Assessment*, *1*(1), 47-65. https://escholarship.org/uc/item/4qp611b4

Osgood, C. E., Suci, G. J., & Tannenbaum, P. (1957). *The Measurement of meaning*. University of Illinois Press.

Piche, G. L., Rubin, D. L., Turner, L. J. & Michlin, M. L. (1978). Teachers' subjective evaluations of standard and Black nonstandard English compositions: A study of written language and attitudes. *Research in the Teaching of English*, *12(2)*, 107-118.

Rubin, D. L., & Williams-James, M. (1997). The impact of writer nationality on mainstream teachers' judgments of composition quality. *Journal of Second Language Writing*, *6*(2), 139-154.

Scharton, M. (1996). The politics of validity. In E. M. White, W. D. Lutz & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 53-75). Modern Language Association.

Schmidt, A. E. & Camara, W. J. (2004). Group differences in standardized test scores and other educational indicators. In R. Zwick (Ed.), *Rethinking the SAT: The future of standardized testing in university admissions* (pp. 189-201). Routledge Falmer.

Shavelson, R. J. (1996). *Statistical reasoning for the behavioral sciences* (3rd ed.). Allyn and Bacon.

Smith, W. L. (1993). Assessing the reliability and adequacy of using holistic scoring of essays as a college composition placement technique. In M. M. Williamson and B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 142-205). Hampton Press.

Smitherman, G. & Villanueva, V. (Eds.). (2003). *Language diversity in the classroom: from intention to practice*. Southern Illinois University Press.

Tabachnick, B. & Fidell, L. S. (2006). *Using multivariate statistics* (5th ed.). Allyn and Bacon.

White, E. M. (2005). The scoring of writing portfolios: Phase 2. *College Composition and Communication*, *56*(4), 581-600.

Williams, F., Whitehead, J. L. & Miller, L. M. (1971). *Attitudinal correlates of children's speech characteristics* (USOE Project No. 0-0336). Center for Communication Research.

Williamson, M. M. & Huot, B. A. (1993). *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*. Hampton Press.

APPENDIX: WRITING PORTFOLIO DIFFERENTIAL SCALE FOR WRITING AND DEMOGRAPHIC INFORMATION

Paper code_____

Circle the appropriate number to indicate your evaluation of the writing.

1. Conception of topic

Unclear	1	2	3	4	5	6	Clear

2. Focus

Unclear	1	2	3	4	5	6	Clear

3. Organization

Disorganized	1	2	3	4	5	6	Organized

4. Support

Not provided	1	2	3	4	5	6	Provided

5. Mechanics

Not Effective	1	2	3	4	5	6	Effective

The writing seems to be

6. Incoherent	1	2	3	4	5	6	Coherent
7. Non-Standard American English	1	2	3	4	5	6	Standard American English
8. Illogical	1	2	3	4	5	6	Logical
9. Ungrammatical	1	2	3	4	5	6	Grammatical
10.Unimaginative	1	2	3	4	5	6	Imaginative
11. Passive	1	2	3	4	5	6	Active
12. Poorly written	1	2	3	4	5	6	Well written

The Student Writer is

13. Weak Writer	1	2	3	4	5	6	Strong Writer
14. Unintelligent	1	2	3	4	5	6	Intelligent
15. Low socio-economic class	1	2	3	4	5	6	High socio-economic class
16. Culturally disadvantaged	1	2	3	4	5	6	Culturally advantaged
17. Unsure	1	2	3	4	5	6	Confident
18. Uncomfortable as a writer	1	2	3	4	5	6	Comfortable as a writer