ePortfolios: Foundational Measurement Issues

by Norbert Elliot, Alex Rudniy, Perry Deess, Andrew Klobucar, Regina Collins, Sharla Sava

Using performance information obtained for program assessment purposes, this quantitative study reports the relationship of ePortfolio trait and holistic scores to specific academic achievement measures for first-year undergraduate students. Attention is given to three evidential categories: consensus and consistency evidence related to reliability/precision; convergent evidence related to validity; and score difference and predictive evidence related to fairness. Interpretative challenges of ePortfolio-based assessments are identified in terms of consistency, convergent, and predictive evidence. Benefits of these assessments include the absence of statistically significant differences in ePortfolio scores for race/ethnicity sub-groups. Discussion emphasizes the need for principled design and contextual information as prerequisite to score interpretation and use. Instrumental value of the study suggests that next-generation ePortfolio-based research must be alert to sample size, design standards, replication issues, measurement of fairness, and reporting transparency. Keywords: ePortfolios, fairness, program assessment, reliability, validity

---

1.0 Introduction

In academic settings, ePortfolios are best defined as digital documentary efforts designed by students to demonstrate proficiencies and record reflections on curricular experiences. Because their design is planned according to institutional outcomes, ePortfolios may then be used by instructors and administrators as locally-developed assessments used to gather evidence of student knowledge and thereby advance learning. In their analysis of 118 peer-reviewed journal articles on ePortfolio research, Bryant and Chittum (2013) found only 18 studies focused on quantitative, outcomes-based research in which student performance was reported. Given the dearth of statistically-informed study findings on ePortfolio outcomes, they concluded that a transition toward empirical assessment is needed. Rhodes, Chen, Watson, & Garrison (2014) agreed, calling for methodological studies yielding evidence of student success and achievement of learning outcomes. In the field of writing assessment, especially needed are three kinds of quantitative studies: reliability studies providing evidence of inter-rater consensus and consistency; studies yielding information about relationships of ePortfolio scores to operationally independent criterion measures; and studies of fairness providing information about sub-group performance. While qualitative studies of ePortfolio-based scores of writing are invaluable for providing contextually rich information, quantitative studies are needed to situate the place of complex assessment genres such as ePortfolios within the ecology of student learning when writing must be scored. Quantitative evidence is especially needed, and often absent, in cases where performance information must be interpreted according to foundational measurement issues involving evidence based on reliability, validity, and fairness.

The primary objective of the present study is therefore to present a principled quantitative study of ePortfolio-based research involving undergraduate post-secondary student writing. We are especially interested in three evidential categories: consensus and consistency evidence related to reliability/precision; convergent evidence related to validity; and score difference and predictive evidence related to fairness. To address questions associated with these categories, we explore relationships between and among ePortfolio holistic scores, ePortfolio trait scores, specific academic performance markers, and demographic factors. These relationships are intended to provide information on the relationship of ePortfolio scores to pre-college measures, enrolled college measures, and predictive college measures. The sample is drawn from an undergraduate student population enrolled in a STEM-oriented technological university located in the northeastern United States. Obtained for program assessment purposes in which writing must be scored, data from this three-year study was analyzed to obtain statistically significant patterns and relationships among defined ePortfolio scores and academic measures. Additionally, post hoc analyses on sub-groups were conducted to evaluate possible differential impact among student sub-groups. Reported are consensus and consistency measures of inter-rater agreement and inter-rater reliability, criterion measures of writing ability external to ePortfolio scores, and description of a model to predict academic performance. Findings from this study add to first generation quantitative research related to score interpretation and use from ePortfolio-based assessments. In terms of lessons learned, the study suggests that next-generation ePortfolio-based research must be alert to sample size, design standards, replication issues, measurement of fairness, and reporting transparency.

2.0 Literature Review

The literature review focuses on three areas: present US and international trends in ePortfolio research; advantages of ePortfolios in writing program assessment; and use of foundational measurement concepts to structure principled analysis. The literature review addresses only concepts related to the present study in terms of trends, program assessment, and foundational measurement categories as these concepts are related to digitally-based assessments; no attempt is made to provide a comprehensive review of the vast literature on portfolio-based assessments. For a comprehensive literature review related to the origin, definitions, rationale, applications, characteristics, and uses of portfolio assessment, see Lam (in press). For a detailed study of the use of portfolios in relationship to faculty development and student learning, see Condon, Iverson, Manduca, Rutz, & Willett (2015).

2.1. ePortfolio Research Trends

Following the United States performance assessment movement of the 1990s, substantial literature has emerged on scoring writing

samples taken under naturalistic conditions (Pullen & Haertel, 2008). Such samples are often classified as performance assessments: tasks requiring students to construct responses over lengthy periods of time, explore alone and with others novel solutions and new task types, and reflect on those responses and solutions. Because of their ability to capture a defined construct, these tasks are best evaluated by disciplinary experts capable of judging advanced knowledge and creative solutions (Lane & Stone, 2006).

In the family of writing performance assessments, Camp (1983) coordinated what is perhaps the earliest United States effort on behalf of the Educational Testing Service. Within a decade, post-secondary programs using portfolios had proliferated to such an extent that Hamp-Lyons and Condon (1993) were able to identify reliability/precision challenges related to multiple text submission, broad use of varied writing tasks, and interpretation of writing process information. Identified also was the need for assessment criteria associated with construct validity, as well as the need to articulate pedagogical impact and build community consensus, two consequential issues related to fairness. In another decade, White (2005) advanced a new vision for portfolio scoring with an emphasis on reflective statements that could be used to examine metacognitive ability.

A milestone in the transition to research on electronic portfolios was the 2003 founding of the National Coalition for Electronic Portfolio Research. In 2006 the organization changed its name to the Inter/National Coalition for Electronic Portfolio Research to reflect new membership from the United Kingdom and Canada (Cambridge, Cambridge, & Yancey, 2009). Internationally, students, teachers, and governing bodies have become the chief stakeholders who stand to benefit, respectively, by pedagogical flexibility, course customized assessment, and adaptability of the ePortfolios in response to diverse curricular demands, as Dysthe, Engelsen, and Lima (2007) have shown in their survey of 69 Norwegian campus leaders. At the present time, research in ePortfolio use may be approached under the categories used by the *International Journal of ePortfolio*, a peer-reviewed journal founded in 2011 with support of the University of Georgia and Virginia Tech: instruction, assessment, and interprofessional/ workplace applications. While theory is absent from these categories, Yancey, McElroy, and Powers (2013) have proposed a new vision for ePortfolio assessment based on personalization, coherence, reflection, judgment, and design.

There is little doubt that a body of knowledge has emerged in ePortfolio-based assessments. Within the United States, the Conference on College Composition and Communication (2015) has developed a position statement, “Principles and Practices in Electronic Portfolios,” that proposes principles, best practices, and learning outcomes to inform ePortfolio use in writing programs. Under the principle of learning outcomes, advice is given to “collaborate with faculty in designing rubrics that consistently facilitate a valid and reliable process of measuring programmatic learning outcomes” (CCCC, 2015). Absent in the extensive bibliography accompanying the position statement is quantitative analysis that could be used to support the premise and practices of the position statement. The present study is therefore intended to address the need for empirical information and to support this expanding body of knowledge.

2.2. ePortfolio-based Program Assessment

Hamp-Lyons (2016) has identified two broad purposes for language assessment: achievement (to measure specific content covered in a course or program) and proficiency (to measure general language command). In US academic settings, assessment of undergraduate student written communication often occurs for the following related purposes: course alignment (e.g., placement of admitted students in the most appropriate course); certification of ability for academic progression and graduation (e.g., exit assessment for a course or matriculation to advanced curricular offerings); review of individual student ability to enhance student learning (e.g., formative classroom assessment); evaluation of curricular effectiveness (e.g., program review); and research (e.g., experimental investigation of student performance) (White, Elliot, & Peckham, 2015). Because the present study is restricted to evaluation of curricular effectiveness through program review, some historical background will be helpful to describe this uniquely US form of assessment.

Program review in the US originated in the early 1990s. As federal debate over educational governance intensified, it became clear that regional agencies would have to demonstrate increasing levels of accountability to ensure that taxpayer dollars were being wisely spent (Parsons, 1997). Assessment of student writing ability (and programs devoted to instruction ensuring that students had basic skills) expanded rapidly when required by the six regional accreditation agencies: the Middle States, the New England Association of Colleges and Schools, the Commission on Institutions of Higher Education, the North Central Association of Colleges and Schools Higher Learning Commission, the Northwest Commission on Colleges and Universities, the Southern Association of Colleges and Schools Commission on Colleges, and the Western Association of Colleges and Schools Accrediting Commission of Senior Colleges and Universities. Prominent programs involving writing placement and progression were instituted in California (White, 2001) and New Jersey (Lutkus, 1985). Today, US program assessment has added advancement of student learning to accountability agendas. Although testing emerged in the twentieth-century as an industry devoted to satisfying accountability demands, advancement of opportunity to learn in the twenty-first century calls for articulated connections between large-scale testing and the instructional environments (Moss, Pullin, Gee, Haertel, & Young, 2008). While there are multiple qualitative and quantitative methods associated with program assessment (Chen, 2015), the present study concentrates only on ePortfolio scores as a form of performance assessment.

Evaluating curricular effectiveness through writing program assessment—the purpose for which the ePortfolio scores were collected in the present study—is best defined as the act validating that those responsible for the program have fulfilled its mission in ways leading to advancement of student learning (While, Elliot, Peckham, 2015; Witte & Faigley, 1983). The New Jersey study site reflects the US history of program assessment. In 1996, first-year writing assessment efforts at the study site shifted from a system of evaluating student best papers with holistic scores (an accountability effort) to a print portfolio system using trait-based methods (an opportunity to learn effort). Portfolios were featured in the 2002 accreditation process as a print-based evaluation. In 2010, the present ePortfolio system began in an effort to include digital forms of writing along with print-based forms (Collins, Elliot, Klobucar, & Deek, 2013). As part of United States program assessment, the study institution employed portfolio assessment to support congruency between the institution and the accreditation requirements established by the Middle States Commission on Higher Education (MSCHE) (Suskie, 2014). Conducted each year, the ePortfolio assessment is congruent with the proposed MSCHE annual updates focusing on assessment of student learning (MSCHE, 2015). ePortfolio assessment is also designed to serve the institution's professional programs that are accredited by discipline-specific accreditation organizations.

2.3. Foundational Measurement Concepts

In terms of measurement foundations, the authors of the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014) provided a common venue for categories of evidence that can be used to design research, interpret results, and plan score use. Combined with the interpretation/use arguments (or IUA) proposed by Kane (2013, 2015) in which validation can be conceptualized as an evaluation of coherence and completeness of the IUA, as well as plausibility of its inferences and assumptions, the *Standards* yields a common referential frame for the present study. While foundational issues of validity, reliability, and fairness will continue to be debated philosophically in both educational measurement and writing studies, categories of evidence established in the *Standards* provide clarity while encouraging meaningful debate (Baird, Hopfenbeck, Newton, Stobart, & Steen-Utheim, 2014; Elliot, 2015; Markus & Borsboom, 2013).

2.3.1. Reliability/Precision: Rater-to-rater evidence. Reliability/precision is defined as score consistency across testing instances of the testing procedure (AERA, APA, NCME, 2014, p. 33). In writing assessment, much of what is known about rater-to-rater evidence (or inter-rater reliability) has been determined under standardized testing conditions seeking to eliminate sources of error rather than situated perspectives in which context reveals the significant activity (Greeno & Gresalfi, 2008). As a result, 0.7 is commonly understood as the lowest acceptable correlation for consistency estimates (Stemler, 2004; Williamson, Xi, and Bryer, 2012), and scores falling below this level are judged as failing to provide evidence of inter-rater reliability. While the relationship between reliability and validity has been debated (Mislevy, 2004; Moss, 2004) and established correlation levels have been challenged (Elliot, Rupp, & Williamson, 2015), additional information is needed to guide score interpretation and use—the interpretation/use argument that includes the "network of inferences and assumptions inherent in the proposed interpretation and use" (Kane, 2015, p. 2)—in complex situated perspectives involving ePortfolio-based assessments.

2.3.2. Validity: Convergent evidence. Validity is defined as the degree to which evidence and theory support interpretation of scores for their proposed uses (AERA, APA, NCME, 2014, p. 11). As a category of evidence, construct representation has received special attention in writing assessment, with locally-developed assessments perceived as yielding robust presence of the writing construct (Condon, 2013). Other evidential categories such as convergent evidence—relationships between ePortfolio scores and other measures intended to assess the same or similar constructs—are less common. Additional information is therefore needed in terms of scores resulting from ePortfolio-based assessments to related measures of academic performances such as admission tests, concurrent course grades, grades in subsequent writing courses, and next semester grade point average.

2.3.3. Fairness: Group difference and differential validity evidence. Fairness is defined as the validity of score interpretation and use for individuals and sub-groups (AERA, APA, NCME, 2014, p. 49). The authors of the *Standards* established that sub-group differences do not, in themselves, indicate lack of fairness but should "trigger follow-up studies" (AERA, APA, NCME, 2014, p. 65). Among such studies, Camilli (2006) noted the importance of general linear modeling techniques in identifying external evidence of bias. Extending the conceptual analysis, Berry (2015) made an important distinction between differential validity and differential prediction in the use of regression modeling. If the correlation with the criterion measures, such as those shown in Table 8, is the same for different subgroups, then they have equal predictive validity even though predictions may be higher for one group. A synonym for predictive bias, differential prediction, conversely suggests a difference in regression lines between subgroups. As addressed in other studies under a program of research dedicated to quantitative study of complex writing samples, this study will address differential validity (Elliot, Deess, Rudniy, & Joshi, 2012; Poe & Cogan, 2016). Because they employ regression techniques and equal risk models, such analyses of differential score function are termed test fair by Hartigan & Wigdor (1989). Generally accepted in the psychometric literature, this form of regression modeling is understood as a "minimum requirement" as defined by Hartigan & Wigdor (1989, p. 255).

However, when turning to classroom assessment—the assessment form most closely related to that of the present study—Camilli (2006) noted that, aside from reliability studies, differential measurement would be "impractical" (p. 246). As the present study

demonstrates, this assumption has not proven accurate. Important information is to be gained from regression analysis (with special attention to the R^2 coefficient of determination) involving ePortfolio assessment. The present study therefore attends to both sub-group differences (statistically significant differences in scores for individuals belonging to sub-groups) and differential validity (under-prediction or over-prediction for individuals within sub-groups).

3.0 Research Questions

The following three research questions were examined in this study. In order to disaggregate information, each question includes analysis of overall and sub-group performance:

1. *Regarding rater-to-rater evidence related to reliability/precision, what levels of consensus and consistency are demonstrated for ePortfolio ratings?* As Table 3 demonstrates, reliability evidence is here provided for the overall student group with ePortfolio scores (N = 210) examined from 2010 to 2012, as well as for male students (n = 179), female students (n = 31), White students (n = 89), Asian students (n = 59), and Hispanic students (n = 36).
2. *Regarding convergent evidence related to validity, what is the relationship of ePortfolio scores to other indicators of academic performance?* As Table 4 illustrates, ePortfolio scores from the overall group and student sub-groups are compared with the overall sample of students to establish validity evidence. For the overall group and for sub-groups, the sampling plan increases when pre-college measures, writing course grades, and predictive college measures are examined in terms of the entire student population. Depending on measure, the overall sampling plan varies: high school rank, n = 1420; SAT Critical Reading, n = 2636; SAT Writing, n = 2636; writing course grade, n = 2171; next writing course grade, n = 2147; next semester GPA, n = 2560). As Tables 5, 6, and 7 illustrate, sub-groups for each of these measures increase accordingly.
3. *Regarding group difference and differential validity evidence related to fairness, what does post-hoc analysis reveal regarding statistically significant differences and coefficients of determination among ePortfolio scores and other indicators of academic performance?* As Tables 1 and 2 demonstrate, evidence regarding group difference varies according to measure. Table 8 includes models of pre-college, enrolled, and predictive measure as they predict ePortfolio holistic score, writing course grade, next writing course grade, and next semester GPA.

4.0 Study Limits

Research identified in 2.1 may be classified as the study of response processes. Such research gathers evidence by questioning instructors about their pedagogical strategies and students about their performance strategies, thus adding to our knowledge about differences in meaning or interpretation of ePortfolio-based assessment. As the *Standards* (AERA, APA, NCME, 2014) noted, theoretical and empirical analyses of response process "can provide evidence concerning the fit between the construct and the detailed nature of the performance or response actually engaged in by the test takers" (p. 15). Substantial knowledge has resulted from studies identified in the literature review that may be classified as validity evidence related to response processes. From questioning students and their instructors to monitoring processes related to task response, evidence related to response processes provides critical information related to interpretation and use. It is important to note that this category of evidence is often absent from information provided about standardized tests.

As to analysis of response processes leading to identification of challenges and benefits to ePortfolio-based scoring, the present study is silent. Post-hoc qualitative empirical analysis regarding the contents of the ePortfolios is not part of this highly restricted study. Employing basic general linear modeling techniques, the present quantitative study is designed exclusively to gather and present information according to categories of evidence identified in the research questions.

5.0 Research Methodology

To understand the study results, some detail is needed to define the situated perspectives of the study site and the methodology used. Such information allows important generalization inferences from the observed sample of performances (as reflected in the ePortfolio scores) to claims about expected performance in the construct sample (as reflected in the criterion measures). While locally developed assessments are often seen as lacking this kind of inference, detailed reporting provides an important bridge from a singular performance to observed performances on multiple measures (Kane, 2013; White, Peckham, & Elliot, 2015). A detailed explication of research methodology also becomes especially important in understanding issues of replication discussed in 8.3.

5.1. Identification of Measures

Twelve measures of student academic ability related to the writing construct are identified in Table 1.

Table 1: ***Measures: Means, Standard Deviations, Range, Statistically Significant Differences, and Effect Sizes***

Admitted First-Time, Full-Time Students					
	Years			Enrollment Percent Increase	
	Fall 2010	Fall 2011	Fall 2012	2010-2011	2011-2012
	N = 948	N = 1016	N = 1045	7.17%	2.85 %
	Measures			Overall Score Difference	
Pre-College Measures					
1. HS Rank	n = 485 *M* = 73 *SD* = 21 Range = 13, 87	n = 453 *M* = 75 *SD* = 20 Range = 12, 100	n = 482 *M* = 74 *SD* = 20 Range = 8, 100	*nss* *t* (963) = -1.4	*nss* *t* (933) = 1.19 2011<2012
2. SAT Critical Reading	n = 857 *M* = 538 *SD* = 83 Range = 320, 800	n = 897 *M* = 536 *SD* = 78 Range = 300, 800	n = 900 *M* = 549 *SD* = 84 Range = 320, 800	*nss* *t* (1734) = .627	2011<2012 *t* (1777) = 3.39** *d* = 0.16
	US N = 1,547,990 *M* = 501 *SD* = 112	NJ n = 84, 847 *M* = 495 *SD* = 113	US N = 1,647,123 *M* = 497 *SD* = 114	NJ n = 86,515 *M* = 495 *SD* = 113	US N = 1,664,479 *M* = 496 *SD* = 114	NJ N = 84,857 *M* = 495 *SD* = 114	2010 NJIT >US *t*(858) = 16.21*** *d* = 0.39 NJIT >NJ *t*(888) = 15.03*** *d* = 0.43	2011 NJIT >US *t*(899) = 14.96*** *d* = 0.40 NJIT >NJ *t*(935) = 15.57*** *d* = 0.42	2012 NJIT >US *t*(901) = 18.91*** *d* = 0.53 NJIT >NJ *t*(901) = 19.27*** *d* = 0.54
3. SAT Writing	n = 857 *M* = 529 *SD* = 87 Range = 280, 800	n = 879 *M* = 533 *SD* = 78 Range = 220, 800	n = 900 *M* = 540 *SD* = 85 Range = 290, 800	*nss* *t*(1734) = -1.05	*nss* *t* (1777) = -1.74
	US N = 1,547,990 *M* = 492 *SD* = 111	NJ n = 84, 847 *M* = 497 *SD* = 115	US N = 1,647,123 *M* = 489 *SD* = 113	NJ n = 86,515 *M* = 497 *SD* = 115	US N = 1,664,479 *M* = 488 *SD* = 114	NJ N = 84,857 *M* = 499 *SD* = 116	2010 NJIT >US *t*(871) = 12.39*** *d* = 0.37 NJIT >NJ t(888) = 10.67*** *d* = 0.31	2011 NJIT >US *t*(899) = 16.89*** *d* = 0.45 NJIT >NJ *t*(936) = 13.67*** *d* = 0.37	2012 NJIT >US *t*(901) = 18.34*** *d* = 0.52 NJIT >NJ *t*(934) = 14.32*** *d* = 0.40
Enrolled College Measures					
4. ePortfolio: Rhetorical Knowledge	n = 41 *M* = 7.41 *SD* = 2.49 Range = 2, 11	n = 50 *M* = 8.06 *SD* = 1.73 Range = 2, 11	n = 119 *M* = 8.29 *SD* = 2.13 Range = 2, 11	*nss* *t* (89) = -1.45	*nss* *t* (167) = -.66
5. ePortfolio: Critical Thinking	n = 41 *M* = 7.05 *SD* = 2.45 Range = 2, 12	n = 50 *M* = 8.12 *SD* = 1.53 Range = 2, 11	n = 119 *M* = 8.06 *SD* = 2.04 Range = 2, 11	2010<2011 *t* (89) = -2.55* d = 0.52	*nss* *t* (167) = .19
6. ePortfolio: Writing Processes	n = 41 *M* = 7.17 *SD* = 2.36 Range = 2, 12	n = 50 *M* = 7.5 *SD* = 1.61 Range = 2, 10	n = 119 *M* = 6.4 *SD* = 1.85 Range = 2, 11	*nss* *t* (89) = -.79	2011>2012 *t* (167) = 3.65*** *d* = 0.63
7. ePortfolio: Knowledge of Conventions	n = 41 *M* = 8.05 *SD* = 2.16 Range = 2, 11	n = 50 *M* = 7.48 *SD* = 1.28 Range = 4, 10	n = 119 *M* = 7.9 *SD* = 2.23 Range = 2, 12	*nss* *t* (89) = .57	*nss* *t* (167) = -.18
8. ePortfolio: Composing in Electronic Environments	n = 41 *M* = 5.95 *SD* = 2.89 Range = 2, 11	n = 50 *M* = 7.62 *SD* = 1.81 Range = 4, 12	n = 119 *M* = 6.34 *SD* = 2.18 Range = 2, 12	2010<2011 *t* (89) = -3.36** *d* = 0.69	2011>2012 *t* (167) = 3.67*** *d* = 0.64

9. ePortfolio: Holistic Score	n = 41 *M* = 7.27 *SD* = 2.79 Range = 2, 12	n = 50 *M* = 8.08 *SD* = 1.66 Range = 2, 10	n = 119 *M* = 7.51 *SD* = 2.11 Range = 2, 12	*nss* *t* (89) = -1.71	*nss* *t* (167) = 1.69
10. Writing Course Grade	n = 680 *M* = 2.91 *SD* = 1.14 Range: 0, 4	n = 725 *M* = 3.08 *SD* = 1.0 Range: 0, 4	n = 767 *M* = 3.01 *SD* = 1.1 Range: 0, 4	2010<2011 *t* (1403) = -3.03*** *d* = .16	*nss* *t* (1409) = 1.27
Predictive College Measures					
11. Next Writing Course Grade	n = 696 *M* = 2.98 *SD* = 1.02 Range: 0, 4	n = 675 *M* = 3.15 *SD* = .89 Range: 0, 4	n = 776 *M* = 3.17 *SD* = .95 Range: 0, 4	2010<2011 *t* (1369) = -3.20*** *d* = .18	*nss* *t* (1449) = -.425
12. Next Semester GPA	n = 823 *M* = 2.76 *SD* = .9 Range = 0, 4	n = 837 *M* = 2.77 *SD* = .88 Range = 0, 4	n = 901 *M* = 2.88 *SD* = .86 Range = 0, 4	*nss* *t* (1658) = -0.4	2011<2012 *t* (1736) = -2.68** *d* = 0.13

Note. *p*-values not statistically significant at the 0.05 level are designated as *nss*.
* $p < .05$
** $p < .01$
*** $p < .001$

Classifications of pre-college measures include high school rank, the SAT Critical Reading Section, and the SAT Writing Section. Enrolled college measures include ePortfolio trait and holistic scores (discussed in 5.5) and grades in the writing class from which the ePortfolio was drawn. Predictive academic measures include grades in the subsequent writing course and next semester Grade Point Average (GPA). Hence, this study reports on relationships among 12 measures identified in Table 1.

In interpreting convergent evidence related to ePortfolio scores, it is understood that the writing construct is most fully represented on a continuum from course grade (robust), ePortfolio scores (targeted), and standardized test scores (narrow). It is also understood that both the SAT Critical Reading and the SAT Writing—admission tests developed by the College Board (Milewski, Johnsen, Glazer, & Kubota, 2005)—are used to reflect a language arts construct combining both reading and writing that is, in turn, reflected in the curriculum of the institution.

As the authors of the *Standards* (AERA, APA, NCME, 2014) noted, identification of measures is of great importance in examining test-criterion relationships. With the exception of the relationship of ePortfolio trait and holistic scores, each of the other pre-college, enrolled, and predictive measures is "operationally distinct" from the ePortfolio "test" (p. 17). Whether the ePortfolio scores are associated with a given measure is therefore a testable hypothesis, one that can lead to further discussion about the relationship of construct representation (a category of validity evidence) to group difference and differential validity (categories of fairness evidence).

5.2. Research Setting

Employing an exploratory, descriptive research design, this study examined an ePortfolio-based assessment used as part of program assessment. The reporting period covers Fall 2010 through Fall 2012, with an update describing present research. The location was a United States public university in New Jersey classified as a Science, Technology, Engineering, and Mathematics dominant research institution by the Carnegie Commission on Higher Education. This institution selectively admits undergraduate and graduate students. A description of admitted students is presented in Table 1.

In pre-college measures, students are in the top quarter of their high schools with statistically significant higher scores on standardized tests than both United States and New Jersey scores. Useful in interpreting statistically significant differences are effect size (*d*) thresholds established by Cohen (1988): small ≤.20; medium ≤.50; and large ≤ .80. Interpretatively, the effect sizes between the SAT Critical Reading and SAT Writing scores of the host institution students, the national sample, and the NJ sample may be generally described as medium—an indication of high performing admitted students. On the other hand, score changes in the SAT Critical Reading between 2011 and 2012 may be characterized as small—an indication of entering student skill stability.

In academic measures, ePortfolio trait and holistic scores (rows 4 through 9 of Table 1) are generally above 7—a score reflecting just-qualified competence—on a 12-point scale in which a score of 12 is the highest and a score of 2 is the lowest. On a 4-point scale, grades (row 10) in the writing course from which ePortfolios are drawn are at the level of B. In predictive measures (rows 11

to 12), subsequent writing course grades are also at that level, and grade point average (GPA) for the overall group is above 2.5.

A university of diverse students, full-time undergraduate enrollment during the period of the study was predominantly male (79%) with a racial/ethnicity profile as follows: White 36%, Asian 22%, Hispanic 21%, Black nine percent, and other sub-groups 12%. Table 2 provides performance measures based on gender assignment and race/ethnicity that reflect equally high levels of performance.

Table 2: ***Total and Disaggregated Measures: Means, Standard Deviations, Range, Statistically Significant Differences, and Effect Sizes***

	Total	Male	Female	Gender Difference	White	Asian	Hispanic	Black	Race/ Ethnicity Difference
Pre-College Measures									
1. HS Rank	N = 1420 *M* =73 *SD* = 21 Range = 8, 100	n = 1155 M_a = 71 *SD* = 21 Range = 8, 100	n = 265 M_a = 80 *SD* = 19 Range = 14, 100	M < F*** *t* (1414) =-6.24 *d* =0.45	n = 502 *M* = 72 *SD* = 21 Range = 8, 100	n = 300 *M* = 75 *SD* = 21 Range = 9, 100	n = 344 *M* = 76 *SD* = 19 Range = 12, 100	n = 154 *M* = 72 *SD* = 19 Range = 9, 99	*nss* *F*(3, 1296) = 2.72
2. SAT Critical Reading	N = 2636 *M* = 540 *SD* = 82 Range = 300, 800	n = 2086 M_a = 537 *SD* = 79 Range = 300, 800	n = 550 M_a = 557 *SD* = 90 Range = 350, 800	M < F*** *t* (2634) =-5.15 *d* = 0.23	n = 974 M_a =559 *SD* = 77 Range = 320, 800	n = 616 M_a = 544 *SD* = 93 Range = 300, 800	n = 510 M_a = 520 *SD* = 73 Range = 340, 760	n = 243 M_a = 519 *SD* = 71 Range = 360, 800	A<W*** *d* = 0.18 H<W*** *d* = 0.52 H<A*** *d* = 0.29 B<W*** *d* = 0.54 B<A*** *d* = 0.30 *F*(3, 2339) = 33.48
3. SAT Writing	N = 2636 *M* = 534 *SD* = 85 Range = 220, 800	n = 2086 M_a = 525 *SD* = 81 Range = 220, 800	n = 550 M_a = 568 *SD* = 94 Range = 310, 800	M < F*** *t* (2634) =-10.58 *d* =0.28	n = 974 M_a =550 *SD* = 76 Range = 340, 800	n = 616 M_a = 550 *SD* = 97 Range = 280, 800	n = 510 M_a = 508 *SD* = 77 Range = 300, 770	n = 243 M_a = 503 *SD* = 75 Range = 350, 750	H<W*** *d* = .0.55 H<A*** *d* = 0.48 B<W*** *d* = 0.62 B<A*** *d* = 0.54 *F*(3, 2339) = 48.39
Enrolled College Measures									
4. ePortfolio: Rhetorical Knowledge	N = 210 *M* = 8.06 *SD* = 2.14 Range = 2, 11	n = 179 M_a = 7.94 *SD* = 2.22 Range = 2, 11	n = 31 M_a = 8.77 *SD* = 1.39 Range = 6, 11	M < F* *t* (208) =-2.03 *d* = 0.49	n = 89 *M* = 8.11 *SD* = 1.97 Range = 2, 11	n = 59 *M* = 8.05 *SD* = 2.03 Range = 2, 11	n = 36 *M* = 8.42 *SD* = 2.01 Range = 3, 11	n = *qns*	*nss* *F*(3, 192) = .81
5. ePortfolio: Critical Thinking	N = 210 *M* = 7.88 *SD* = 2.06 Range = 2, 12	n = 179 M_a = 7.73 *SD* = 2.09 Range = 2, 12	n = 31 M_a = 8.74 *SD* = 1.57 Range = 5, 12	M < F** *t* (208) =-2.57 *d* = .055	n = 89 *M* = 7.80 *SD* = 1.94 Range = 2, 11	n = 59 *M* = 8.03 *SD* = 1.90 Range = 2, 12	n = 36 *M* = 8.31 *SD* = 1.93 Range = 3, 12	n = *qns*	*nss* *F*(3, 192) = 1.59

6. ePortfolio: Writing Processes	N = 210 *M* = 6.81 *SD* = 1.96 Range = 2, 12	n = 179 *M* = 6.6$_a$ *SD* = 2.01 Range = 2, 12	n = 31 *M* = 7.71$_a$ *SD* = 1.37 Range = 6, 11	M < F** *t* (208) =-2.80 *d* = 0.64	n = 89 *M* = 6.62 *SD* = 1.93 Range = 2, 11	n = 59 *M* =7.15 *SD* = 1.93 Range = 2, 12	n = 36 *M* = 7.06 *SD* = 1.84 Range = 4, 10	n = *qns*	*nss* *F*(3, 192) = 2.69
7. ePortfolio: Knowledge of Conventions	N = 210 *M* = 7.91 *SD* = 2.02 Range = 2, 12	n = 179 *M* = 7.79$_a$ *SD* = 2.05 Range = 2, 12	n = 31 *M* = 8.65$_a$ *SD* = 1.74 Range = 4, 12	M < F* *t* (208) =-2.19 *d* = 0.45	n = 89 *M* = 7.96 *SD* = 1.88 Range = 2, 12	n = 59 *M* = 8.05 *SD* = 1.98 Range = 2, 12	n = 36 *M* = 8.22 *SD* = 1.59 Range = 4, 11	n = *qns*	*nss* *F*(3, 192) = 1.78
8. ePortfolio: Composing in Electronic Environments	N = 210 *M* = 6.57 *SD* = 2.33 Range = 2, 12	n = 179 *M* = 6.45 *SD* = 2.36 Range = 2, 12	n = 31 *M* = 7.26 *SD* = 2.02 Range = 4, 12	*nss*	n = 89 *M* = 6.33 *SD* = 2.27 Range = 2, 12	n = 59 *M* = 6.69 *SD* = 2.13 Range = 2, 12	n = 36 *M* = 6.86 *SD* = 2.36 Range = 2, 12	n = *qns*	*nss* *F*(3, 192) = .62
9. ePortfolio: Holistic Score	N = 210 *M* = 7.60 *SD* = 2.17 Range = 2, 12	n = 179 *M* = 7.46$_a$ *SD* = 2.21 Range = 2, 11	n = 31 *M* = 8.39$_a$ *SD* = 1.76 Range = 5, 12	M < F* *t* (208) =-2.20 *d* = 0.50	n = 89 *M* = 7.58 *SD* = 1.98 Range = 2, 12	n = 59 *M* = 7.71 *SD* = 1.94 Range = 2, 11	n = 36 *M* = 8.06 *SD* = 2.27 Range = 2, 12	n = *qns*	*nss* *F*(3, 192) = 1.83
10. Writing Course Grade	N = 2172 *M* = 3.0 *SD* = 1.09 Range = 0, 4	n = 1727 *M* = 2.94$_a$ *SD* = 1.11 Range = 0, 4	n = 444 *M* = 3.24$_a$ *SD* = .96 Range = 0, 4	M < F*** *t* (2169) =-5.24 *d* = 0.29	n = 856 *M* = 3.11$_a$ *SD* = 1.08 Range = 0, 4	n = 498 *M* = 3.1$_a$ *SD* = 1.0 Range = 0, 4	n = 391 *M* = 2.87$_a$ *SD* = 1.06 Range = 0, 4	n = 199 *M* = 2.75$_a$ *SD* = 1.14 Range = 0, 4	H<W** *d* = 0.22 H<A** *d* = 0.22 B<W*** *d* = 0.32 B<A*** *d* = 0.33 *F*(3, 1940) = 10.16
Predictive College Measures 11.Next Writing Course Grade	N = 2147 *M* = 3.11 *SD* = .96 Range = 0, 4	n = 1678 *M* = 3.04 *SD* = .98 Range = 0, 4	n = 469 *M* = 3.34 *SD* = .93 Range = 0, 4	M < F*** *t* (2145) =-6.05 *d* = 0.31	n = 810 *M* = 3.24 *SD* = .89 Range = 0, 4	n = 517 *M* = 3.13 *SD* = .94 Range = 0, 4	n = 403 *M* = 3.01 *SD* = .97 Range = 0, 4	n = 201 *M* = 2.81 *SD* = 1.11 Range = 0, 4	H<W** *d* = 0.25 B<W*** *d* = 0.43 B<A*** *d* = 0.31 *F*(3,1927) = 13.17
12. Next Semester GPA	N = 2560 *M* = 2.8 *SD* = .88 Range = 0, 4	n = 2013 *M* = 2.76$_a$ *SD* = .89 Range = 0, 4	n = 547 *M* = 2.98$_a$ *SD* = .837 Range = 0, 4	M < F*** *t* (2169) =-5.41 *d* = 0.25	n = 937 *M* = 2.9$_a$ *SD* = .85 Range = 0, 4	n = 620 *M* = 2.88$_a$ *SD* = .88 Range = 0, 4	n = 485 *M* = 2.73$_a$ *SD* = .83 Range = 0, 4	n = 243 *M* = 2.51$_a$ *SD* = .89 Range = 0, 4	H<W** *d* = 0.20 H<A* *d* = 0.17 B<W*** *d* = 0.49 B<A*** *d* = 0.42 B<H** *d* = 0.26 *F*(3, 2281) = 16.13

Note. Different subscripts ($_a$) within a row represent means different by independent sample t-test (2-tailed) for gender and by Bonferroni correction for race/ethnicity. Sample sizes under 30, too small for inferential analysis, are designated *qns* (quantity not sufficient). *P*-values not statistically significant at the 0.05 level are designated as *nss*.
* $p < .05$
** $p < .01$
*** $p < .001$

Scores on standardized tests are higher for all sub-groups at the study site than national scores disaggregated by gender and race/ethnicity (College Board, 2012, Table 9). Present and next semester writing course grades, as well as GPA, are at the level of C+ or above for all sub-groups.

5.3. Sampling Plan

The portrait of admitted students shown in Table 1 excludes students who used ACT scores or others who were admitted under special programs. Because the target population of the admitted class included approximately 1,000 first-time, full-time first years students, a sampling plan was developed to obtain a representative sample allowing generalizability inferences from the sample to the target population. Because it is time consuming to trait score, a sampling plan was implemented to identify the smallest possible number of ePortfolios to be assessed (Bridgeman, Ramineni, Dean, & Li, 2014; White, Elliot, & Peckham, 2015). After the target number of ePortfolios was identified for each year of the study, samples were randomly selected across sections of the first semester writing course. With confidence intervals between 0.80 and 0.95 depending on year, the sampling plan resulted in sub-group sizes reflective of the overall undergraduate population (see Table 2).

While 12 ePortfolios were collected for Black students (underrepresented by three percent less), that total fell below the minimum number acceptable for inferential statistical analysis. Specifically, we follow the time-honored rule of thumb offered by Roscoe (1969) in advising that sample sizes smaller than 30 are to be avoided because they cannot assure the benefits of the central limit theorem (pp. 156-157).

Justification for combining the three years of the study is based on consistency of admitted students and absence of statistically

significant differences on the majority of ePortfolio trait and holistic scores (see Table 1). The sample consisted of ePortfolios from 210 students (males: n = 179; females: n = 31; White: n = 89; Asian: n = 59; Hispanic: n = 36; Black: n = 26). The only statistically significant difference that proved constant across the three years of the study was found in ePortfolio trait scores on composing in electronic environments, discussed in 7.3. Because of score consistency over time, combining scores over a three-year period was justified. The aggregated scores across three years allowed a larger sampling plan for the total group and yielded additional sub-group analysis.

As noted in the three research questions, different sample sizes are used for different analyses. Available data is provided for a given cohort or by combining cohorts. For instance, depending on measure, all admitted students do not have all information: some high schools do not supply rank; and some students take the ACT test instead of the SAT. Because ePortfolios were required of all undergraduate students by university policy during the time of the study, we have been careful to use all common available data and can identify little or no overall participant attrition in the collection of ePortfolio scores during the study period—a reporting guideline associated with transparency and supported by the Institute of Education Sciences (2014).

A final justification of the sampling plan is needed in terms of using larger student samples (e.g., SAT Writing = 2636) to determine relationships with our ePortfolio scores (e.g., ePortfolio holistic score = 210). In Tables 1 and 2, for instance, readers may wonder why SAT Critical Reading scores (N = 2636) was compared to ePortfolio holistic scores (n = 210). The following two analyses provide justification for using both samples. Both analyses are based on information provided in Tables 1 and 2.

First, comparison between the overall sample and the ePortfolio sample revealed absence of statistical significance on five of the six criterion measures: HS Rank ($t(138) = .54$, $p = .59$); SAT Critical Reading ($t(258) = 1.94$, $p = .06$); SAT Writing ($t(135) = 1.88$, $p = .06$); writing course grade ($t(232) = 1.53$, $p = .13$); and next writing course grade ($t(222) = 0$, $p = .10$). Statistically significant values were, however, observed regarding next semester GPA: ($t(256) = 3.07$, $p < .001$, $d = .20$). With only one statistically significant difference at a small effect size, it was clear that there were advantages to using the larger comparative sample.

At the next stage of analysis, comparison of sub-group means was used to determine if statistically significant differences in the larger overall sample were different from those in the smaller ePortfolio sample. In the case of the ePortfolios, only one statistically significant difference was observed: SAT Writing ($F(1, 200) = .39$, $p < .05$). For each of the other criterion measures, there were no statistically significant gender differences for students with ePortfolio scores: SAT Verbal ($F(1, 200) = .39$, $p = .53$); HS rank ($F(1, 113) = 2.08$, $p = .15$); writing course grade ($F(1, 205) = .49$, $p = .49$); next writing courses grade ($F(1, 185) = 3.52$, $p = .06$); and next semester GPA ($F(1, 201) = .01$, $p = .91$). In an analysis of ethnicity, no statistically significant differences were reported for any sub-groups of student with ePortfolio scores: SAT Verbal ($F(7, 194) = 1.80$, $p = .09$); SAT Writing ($F(7, 194) = .97$, $p = .45$); HS rank ($F(7, 107) = .48$, $p = .85$); writing course grade ($F(7, 199) = 1.01$, $p = .43$); next writing course grade ($F(7, 179) = .81$, $p = .58$); and next semester GPA ($F(7, 195) = .47$ $p = .86$). This lack of statistically significant difference conflicted with the many statistically significant differences in criterion measures illustrated in Table 2 sub-group comparisons. Were we to have used only the smaller sample, these important sub-group differences of the larger sample would have been masked. In order to preserve emphasis on evidence related to fairness by emphasizing the need to record sub-group differences, we therefore decided to use the different sample sizes shown in Tables 1 and 2 in the present study.

5.4. ePortfolio Program

Consisting of a two-course sequence, the curriculum is taught by a core cohort of full-time instructors who both teach and assess students in the first-year curriculum. When adjuncts are employed, they are mentored by the director of composition. All instructors involved in portfolio assessment have in-depth experience in both trait and holistic assessment and may be considered a community.

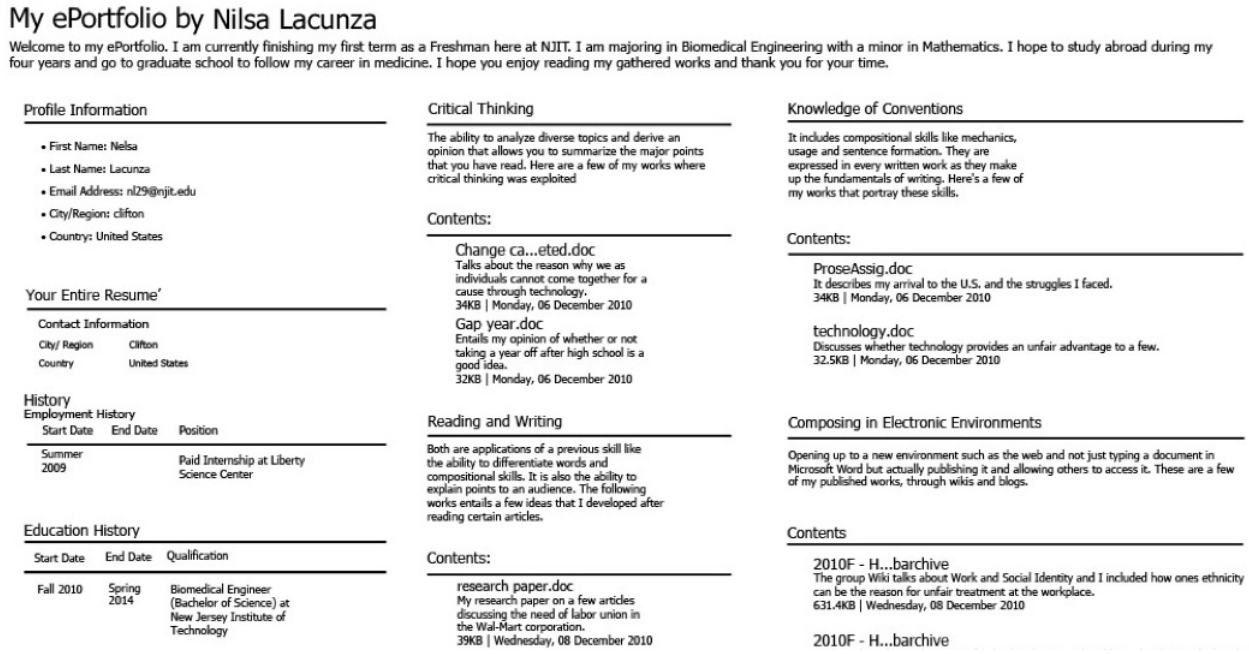
This study captures the ePortfolio scores of students in first-year writing during a period of instructional transition from print-based writing to digitally-based writing. While the curriculum retained source-based writing, new interactive elements—blogs, podcasts, and collaborative wikis—were introduced beginning in 2010. As well, ePortfolios were also introduced with the new digital genres, and instructors were encouraged to work with students so their best work was presented for scoring. During the Fall 2010 stage of the study, the assessment reflected national standards of best practice derived from the WPA Outcomes Statement—a US consensus view intended to introduce first-year postsecondary students to writing expectations (Council of Writing Program Administrators, 2000, 2008; Dryer, Bowden, Brunk-Chavez, Harrington, Halbritter, & Yancey, 2014). In practice, the five writing, reading, and critical analysis experiences of the WPA Outcomes Statement became both course objectives and trait scoring criteria: rhetorical knowledge; critical thinking, reading, and writing; processes; knowledge of conventions; and composing in electronic environments. While paper portfolios had been scored according to traits of rhetorical knowledge, critical thinking, writing processes, and knowledge of conventions associated with a holistic score, during the 2010 operational phase of ePortfolio implementation, a fifth trait—composing in electronic environments—was added to capture experiences associated with digital forms of writing.

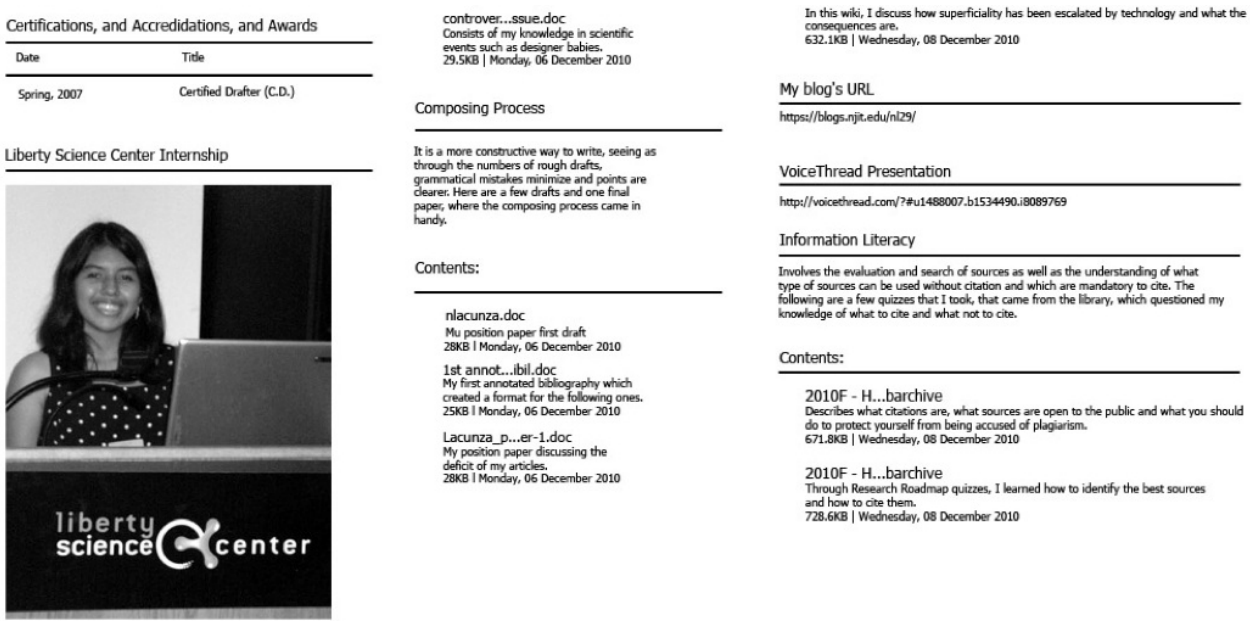
The origin, development, limits, and usefulness of the WPA Outcomes Statement are well established and widely recorded (Behm, Glau, Holdstein, Roen, & White, 2013; Harrington, Rhodes, Fischer, & Malencyzk, 2005; Isaacs & Knight, 2013). Relevant to the

present study are three advantages obtained when the WPA Outcomes Statement is used in writing program design and assessment. First, when used as both course objectives and scoring rubric, instructors leverage the five outcomes to achieve instructional consistency that, in turn, fosters inter-rater reliability during assessment episodes. When ePortfolios are assembled and assessed according to well-articulated criteria, diverse forms of student work become scoreable in a more predictable fashion because the circle of instruction and assessment is complete (Panadero & Jonsson, 2013). Indeed, the ePortfolio itself can be considered as a form of constructed response task (Bennett, 1993; White, Elliot, & Peckham, 2015). Second, the five outcomes may also be understood as variables of the writing construct. These variables allow instructors to model the writing construct and explore how it shifts according to genre and audience; in turn, these instructional variables lend construct validity to the trait scores used in the assessment (Kelly-Riley & Elliot, 2014). Third, the transparency and detail of the WPA Outcomes Statement provide affordances for students that include access to resources and practices associated with fairness and opportunity to learn (Greeno & Gresalfi, 2008). The writing construct thus becomes situated, allowing students to participate in writing instruction in meaningful ways because course outcomes, variable modeling, and scoring criteria are uniformly and consistently presented.

Regarding that which was scored, program administrators and instructors considered the ePortfolio a targeted representation of the writing construct. That is, the ePortfolio was intended to capture student writing performance longitudinally over the period of a semester. Students were asked to upload their best work as evidence of their proficiency with the five outcomes, along with a brief reflective statement accompanying each outcome. A model ePortfolio is shown in Figure 1.

Figure 1: ePortfolio of Nilsa Lacunza, NJIT first-year student.
(Originally published in White, Peckham, & Elliot, 2015. Used by permission of Ms. Lacunza and Utah State University Press.)





As Figure 1 illustrates, students were invited to create five sections of the ePortfolio, each representing a distinct trait. Files and links representing both traditional print-based and innovative digital assignments were then uploaded to each section, accompanied by a brief reflective statement intended to promote metacognition associated with knowledge transfer (National Research Council, 2012; Yancey, Robertson, & Taczak, 2014). The ePortfolio was thus intended to allow students an opportunity to demonstrate their experiences with writing as a process, to acknowledge that audiences who view the digital artifact extend beyond the instructor, and to reflect on their experiences with those varied assignments and multiple audiences.

5.5. Scoring Procedures

ePortfolios received scores on each of the five traits and a holistic score. Based on classifications by Stemler (2004), information is presented on consensus estimates (based on the assumption that raters should be able to come to exact or adjacent agreement about how to apply levels of a scoring rubric to the observed behaviors) and consistency estimates (based upon the assumption that each judge is consistent in classifying the levels of the rubric). Consensus estimates are presented in percent-agreement calculations of exact agreement (e.g., 6+6) and adjacent scores (e.g., 6+5). Scores differing by 2 (e.g., 6+4), 3 (e.g., 6+3), and 4 points (e.g., 6+2) are considered discrepant and adjudicated by a third rater. Hence, if rhetorical knowledge is awarded a trait score of 6 + 4, that trait is scored by a third rater whose judgment determines the final score; if the third rater awards a score of 5, the total final score is 11. Consistency estimates are presented using both Pearson correlations (*r*) and quadratic weighted kappa correlations (*K*). Estimates in Table 3 are provided for all students, as well as for certain sub-groups—males, females, White students, Asian students, and Hispanic students. Because of insufficient sample size, ePortfolio scores are not provided for Black students.

Scores were recorded using a web-based portfolio assessment application (WebPAA) described by Collins, Elliot, Klobucar, & Deek (2013). Relevant to the present study is that the scoring protocols allow accurate, real-time score recording and discrepancy resolution. With 10 raters using the system, 50 ePortfolios can be awarded trait and holistic scores, with discrepancies resolved, in a single day.

Uniform calibration sessions used by the university were used to score the 2010 ePortfolios in the study (Elliot, Briller, & Joshi, 2007). Program administrators selected ePortfolios evidencing ranges on the 6-point Likert scale for each trait and for the holistic score. Following an orientation and training session, raters subsequently scored and discussed URLs linked to ePortfolios and raters synchronously scored student work using the WebPAA system. To accommodate increased sample size for 2012, ePortfolios were scored asynchronously. ePortfolios were read immediately after final grades were due, thus preventing students from removing their ePortfolios from the database. As well, the timing of the reading assured that writing course grades were operationally distinct from ePortfolios trait and holistic scores, as discussed in 5.1.

Our methods therefore produced information on inter-rater reliability. Resource allocation did not allow us to produce information on the reliability of the adjudicated scores that are actually used in the analysis, a process that would require correlation of scores from two independent adjudications of a set of scores. Focus should therefore be on the correlations with external variables because those correlations suggest some level of reliability in the adjudicated scores.

5.6. Measures

To address Research Question 1, we used simple counts to establish consensus. To establish consistency estimates, we used both Pearson product moment correlation coefficients and weighted Kappa coefficients. As Table 3 demonstrates, use of both measures allows reporting transparency. Following White, Elliot, & Peckham (2015), interpretative ranges are provided in 7.1. To address Research Question 2, we used Pearson product moment correlations coefficients. Tables 4, 5, 6, and 7 illustrate the usefulness of this basic method for the overall group and sub-groups. Following Kelly-Riley & Elliot (2016), interpretive ranges are provided in 6.2. As shown in Table 1, Study 3 used 2-tailed tests of significance to identify score differences in sub-groups. Sub-group differences were further analyzed using the Bonferroni correction, as demonstrated in Table 2. Study 3 also employed linear regression techniques to investigate relationships among predictor and outcome variables. Because data transparency was a goal of the study, no attempt was made to correct for model over-fitting. When effect size is reported in Tables 1, 2, and 8, we have used Cohen's *d* (1992), as calculated by Ellis (2009). Interpretive ranges are provided in 5.2. For all studies, statistical significance is reported at 0.05, 0.01, and 0.001. Unless $p < .05$ the result is reported as not statistically significant (*nss*). Basic descriptive and inferential measures are useful in demonstrating the power of quantitative analysis for complex assessments using ePortfolios. As well, because the techniques are generally used, replicability of evidence can be gathered across settings, as discussed in 8.3.

5.7. Software

Information provided in Tables 1, 2, 4, 5, 6, 7, and 8 was analyzed in SPSS 20. Information provided in Table 3 was analyzed in R 3.2.3 using the Design package.

6.0 Results

Results are categorized in terms of evidence related to reliability, validity, and fairness. Based on the study design, we use a variety of descriptive and inferential tables to present ePortfolio scores as they are related to each other and to other measures of the writing construct. Because these results are first generation quantitative research involving ePortfolio trait scores, we have disaggregated data to deepen analysis.

6.1 Reliability: Consensus and Consistency Evidence

In terms of inter-rater consensus and consistency estimates, Table 3 presents a detailed portrait of the challenges involved in scoring ePortfolios for the overall sample and for sub-groups. Presented are consensus estimates of exact, adjacent, and discrepant scores. Presented also are two types of consistency estimates: Pearson and weighted Kappa coefficients. Reporting practices follow those recommended by Hallgren (2012) for observational ratings provided by multiple readers. The 0.7 coefficient is used as the lowest acceptable correlation for consistency estimates; however, this standard gauge is problematized in 7.1.

Table 3: *ePortfolio Consensus and Consistency Estimates, All Groups*

	Consensus Estimates					Consistency Estimates			
	Exact agreement	Adjacent	Scores differ by 2	Scores differ by 3	Scored differ by 4	Non-Adjudicated Pearson	Adjudicated Pearson	Non-Adjudicated Weighted Kappa	Adjudicated Weighted Kappa
All Students (N = 210)									
ePortfolio: Rhetorical Knowledge	73 (35%)	87 (41%)	41 (20%)	7 (3%)	2 (1%)	.42***	.67***	.41***	.67***
ePortfolio: Critical Thinking	88 (42%)	88 (42%)	28 (13%)	6 (3%)	0 (0%)	.54***	.71***	.53***	.70***
ePortfolio: Writing Processes	79 (38%)	81 (39%)	40 (19%)	8 (4%)	2 (1%)	.37***	.59***	.37***	.58***
ePortfolio: Knowledge of Conventions	71 (34%)	86 (41%)	47 (22%)	6 (3%)	0 (0%)	.43***	.67***	.41***	.66***
ePortfolio: Composing in Electronic Environments	88 (42%)	77 (27%)	31 (15%)	12 (6%)	2 (1%)	.53***	.76***	.52***	.75***
ePortfolio: Holistic Score	83 (40%)	92 (44%)	27 (13%)	7 (3%)	1 (0%)	.53***	.77***	.51***	.77***
Male Students (n = 179)									
ePortfolio: Rhetorical Knowledge	65 (36%)	68 (38%)	37 (21%)	7 (4%)	2 (1%)	.44***	.69***	.42***	.69***
ePortfolio: Critical Thinking	75 (42%)	75 (42%)	24 (13%)	5 (3%)	0 (0%)	.56***	.71***	.55***	.71***
ePortfolio: Writing Processes	68 (38%)	70 (39%)	33 (18%)	7 (4%)	1 (1%)	.40***	.64***	.40***	.64***
ePortfolio: Knowledge of Conventions	62 (35%)	73 (41%)	38 (21%)	6 (3%)	0 (0%)	.44***	.42***	.67***	.67***
ePortfolio: Composing in Electronic Environments	81 (45%)	62 (35%)	25 (14%)	10 (6%)	1 (1%)	.57***	.56***	.80***	.79***
ePortfolio: Holistic Score	70 (39%)	79 (44%)	22 (12%)	7 (4%)	1 (1%)	.53***	.51***	.78***	.78***
Female Students (n = 31)									
ePortfolio: Rhetorical Knowledge	8 (26%)	19 (61%)	4 (13%)	0 (0%)	0 (0%)	.19 ns	.35*	.19 ns	.35*
ePortfolio: Critical Thinking	13 (42%)	13 (42%)	4 (13%)	1 (3%)	0 (0%)	.29 ns	.54**	.28 ns	.52**
ePortfolio: Writing Processes	11 (35%)	11 (35%)	7 (23%)	1 (3%)	1 (3%)	-.04 ns	.02 ns	-.04 ns	.02 ns
ePortfolio: Knowledge of Conventions	9 (29%)	13 (42%)	9 (29%)	0 (0%)	0 (0%)	.18 ns	.58***	.18 ns	.58**
ePortfolio: Composing in Electronic Environments	7 (23%)	15 (48%)	6 (19%)	2 (6%)	1 (3%)	.22 ns	.48**	.22 ns	.48**
ePortfolio: Holistic Score	13 (42%)	13 (42%)	5 (16%)	0 (0%)	0 (0%)	.44**	.63***	.44**	.62***
White Students (n = 89)									
ePortfolio: Rhetorical Knowledge	32 (36%)	39 (44%)	15 (17%)	3 (3%)	0 (0%)	.49***	.67***	.47***	.67***
ePortfolio: Critical Thinking	41 (46%)	33 (37%)	12 (13%)	3 (3%)	0 (0%)	.55***	.73***	.52***	.73***
ePortfolio: Writing Processes	35 (39%)	35 (39%)	14 (16%)	5 (6%)	0 (0%)	.40***	.62***	.40***	.62***
ePortfolio: Knowledge of Conventions	33 (37%)	35 (39%)	19 (21%)	2 (2%)	0 (0%)	.49***	.67***	.45***	.67***
ePortfolio: Composing in Electronic Environments	44 (49%)	30 (34%)	13 (15%)	2 (2%)	0 (0%)	.63***	.83***	.61***	.81***
ePortfolio: Holistic Score	36 (40%)	39 (44%)	12 (13%)	2 (2%)	0 (0%)	.57***	.77***	.53***	.76***

Asian Students (n = 59)									
ePortfolio: Rhetorical Knowledge	20 (34%)	25 (42%)	12 (20%)	1 (2%)	1 (2%)	.33**	.61***	.33**	.60***
ePortfolio: Critical Thinking	25 (42%)	27 (46%)	5 (8%)	2 (3%)	0 (0%)	.52***	.51***	.62***	62***
ePortfolio: Writing Processes	20 (34%)	25 (42%)	12 (20%)	2 (3%)	0 (0%)	.35**	.37**	.56***	.56***
ePortfolio: Knowledge of Conventions	17 (29%)	28 (47%)	13 (22%)	1 (2%)	0 (0%)	.34**	.34**	.67***	.66***
ePortfolio: Composing in Electronic Environments	20 (34%)	29 (49%)	6 (10%)	3 (5%)	1 (2%)	.50***	.42***	.64***	.63***
ePortfolio: Holistic Score	23 (39%)	29 (49%)	6 (10%)	1 (2%)	0 (0%)	.53***	.74***	.52***	.72***
Hispanic Students (n = 36)									
ePortfolio: Rhetorical Knowledge	10 (28%)	15 (42%)	9 (25%)	2 (6%)	0 (0%)	.16nss	.61***	.16nss	.60***
ePortfolio: Critical Thinking	11 (31%)	17 (47%)	8 (22%)	0 (0%)	0 (0%)	.36*	.60***	.35*	.59***
ePortfolio: Writing Processes	13 (36%)	14 (39%)	8 (22%)	1 (3%)	0 (0%)	.39*	.67***	.36*	.64***
ePortfolio: Knowledge of Conventions	11 (31%)	14 (39%)	10 (28%)	1 (3%)	0 (0%)	.15nss	.34*	.15nss	.33*
ePortfolio: Composing in Electronic Environments	12 (33%)	10 (28%)	9 (25%)	4 (11%)	0 (0%)	.24nss	.66***	.24nss	.66***
ePortfolio: Holistic Score	12 (33%)	15 (42%)	6 (17%)	2 (6%)	0 (0%)	.28nss	.73***	.28nss	.73***

* $p < .05$
** $p < .01$
*** $p < .001$
Note: p-values not statistically significant at the 0.05 level are designated as *nss*.

As Table 3 shows, only adjudicated scores for critical thinking, composing in electronic environments, and the holistic score from the overall group would be admissible for further interpretation and use. No inferences whatsoever could be made for any of the scores for female students, and only the holistic scores could be used for Hispanic students. Put another way, few of the correlations reported in Table 3 would be seen as acceptable for further study. Furthermore, if we recall the observation of Haertel (2006) that adjudication violates statistical assumptions, then only a single non-adjudicated weighted kappa score from Table 3 ($K = 0.80$, $p < .001$, for male students on composing in electronic environments) met the 0.7 standard. Were the 0.7 coefficient used, only a single score on a single trait for a single sub-group could be used for further interpretation and use.

6.2. Validity: Convergent Evidence

Tables 4 to 7 present correlations relating to pre-admission measures, ePortfolio traits, academic measures, and predictive measures for the overall population and for all sub-groups. The correlation ranges used in analyses and discussions are as follows: high positive correlations = 1.0 to 0.70; medium positive correlations = 0.69 to 0.30; and low positive correlations = 0.29 to 0.00.

Table 4: ***Correlation of Measures: All Students***

Measures	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
1. HS Rank (N = 1420)	—	.21**	.30**	.12	.14	.27**	.23**	.23*	.20*	.33**	.32**	.40**
2. SAT Critical Reading (N = 2636)		—	.73**	.11	.05	.07	.18**	-.00	.14*	.16**	.24**	.26**
3. SAT Writing (N = 2636)			—	.16*	.13	.13	.29**	.03	.16*	.24**	.27**	.33**
4. ePortfolio: Rhetorical Knowledge (N = 210)				—	.84**	.60**	.71**	.59**	.84**	.10	.18*	.17**
5. ePortfolio: Critical Thinking (N = 210)					—	.62**	.71**	.57**	.82**	.14*	.26**	.17*
6. ePortfolio: Writing Processes (N = 210)						—	.61**	.50**	.70**	.14*	.24**	.20**
7. ePortfolio: Knowledge of Conventions (N = 210)							—	.44**	.73**	.19**	.21**	.17*
8. ePortfolio: Composing in Electronic Environments (N = 210)								—	.69**	.05	.12	.18*
9. ePortfolio: Holistic Score (N = 210)									—	.18*	.20**	.25**
10. Writing Course Grade (N = 2171)										—	.39**	.70**
11. Next Writing Course Grade (N = 2147)											—	.45**
12. Next Semester GPA (N = 2560)												—

*$p < .05$
**$p < .01$

Table 5: ***Correlation of Measures: Male and Female Students***

Measures	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
1. HS Rank (Men = 1155) (Women = 265)	Men→ Women↓	.18**	.25**	.08	.09	.22*	.20*	.19	.14	.30**	.31**	.37**
2. SAT Critical Reading (Men = 2086) (Women = 550)	.36**	Men↑ — Women↓	.713**	.10	.05	.06	.19*	-.01	.14	.13**	.23**	.23**
3. SAT Writing (Men= 2086) (Women = 550)	.41**	.78**	Men↑ — Women↓	.13	.11	.07	.27**	-.01	.12	.22**	.24**	.31**
4. ePortfolio: Rhetorical Knowledge (Men= 179) (Women = 31)	.508	.16	.24	Men↑ — Women↓	.85**	.58**	.71**	.58**	.85**	.10	.18**	.18**
5. ePortfolio: Critical Thinking (Men = 179) (Women = 31)	.51	-.04	.06	.71**	Men↑ — Women↓	.62**	.72**	.57**	.82**	.14	.26**	.18**
6. ePortfolio: Writing Processes (Men = 179) (Women = 31)	.77**	.04	.38*	.65**	.54**	Men↑ — Women↓	.60**	.48**	.69**	.15*	.24**	.21**
7. ePortfolio: Knowledge of Conventions (Men = 179) (Women = 31)	.58**	.12	.30	.59**	.55**	.54**	Men↑ — Women↓	.46**	.74**	.20**	.21**	.19*
8. ePortfolio: Composing in Electronic Environments (Men = 179) (Women = 31)	.59**	.00	.13	.51**	.53**	.50**	.21	Men↑ — Women↓	.70**	.04	.16**	.18*
9. ePortfolio: Holistic Score (Men = 179) (Women = 31)	.74**	.07	.25	.69**	.76**	.75**	.61**	.58**	Men↑ — Women↓	.19*	.19**	.26**
10. Course Grade (Men = 1727) (Women = 444)	.47**	.28**	.28**	.20	.23	.23	.19	.14	.19	Men↑ — Women↓	.39**	.7**
11. Next Writing Course Grade (Men = 1678) (Women = 469)	.32**	.26**	.31**	-.01	-.07	.02	.05	-.47**	.02	.38**	Men↑ — Women↓	.44**
12. Next Semester GPA (Men = 2013) (Women = 547)	.54**	.33**	.35**	.15	.20	.20	.11	.21	.23	.68**	.44**	Men↑ ←Women

*$p < .05$
**$p < .01$

Table 6: ***Correlation of Measures: White and Asian Students***

Measures	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12
1. HS Rank (White = 502) (Asian = 300)	White→ Asian↓	.28**	.38**	.04	.06	.17	.16	.19	.06	.36**	.38**	.45**
2. SAT Critical Reading (White = 974) (Asian = 616)	.31**	White↑ — Asian↓	.70**	.07	-.02	.09	.18	.06	.17	.16**	.22**	.27**
3. SAT Writing (White = 974) (Asian = 616)	.37**	.78**	White↑ — Asian↓	.09	.11	.02	.25*	.03	.12	.25**	.25**	.34**
4. ePortfolio: Rhetorical Knowledge (White = 89) (Asian = 59)	.19	.23	.41*	White↑ — Asian↓	.79**	.54**	.67**	.54**	.85**	.24*	.12	.17
5. ePortfolio: Critical Thinking (White = 89) (Asian = 59)	.23	.14	.32*	.84**	White↑ — Asian↓	.57**	.65**	.46**	.79**	.30**	.25*	.19
6. ePortfolio: Writing Processes (White = 89) (Asian = 59)	.31	.08	.45**	.58**	.63**	White↑ — Asian↓	.55**	.45**	.64**	.32**	.23*	.28**
7. ePortfolio: Knowledge of Conventions (White = 89) (Asian = 59)	.31	.23	.41*	.67**	.72**	.62**	White↑ — Asian↓	.39**	.71**	.38**	.23*	.20
8. ePortfolio: Composing in Electronic Environments (White = 89) (Asian = 59)	.07	-.02	.16	.54**	.62**	.54**	.40**	White↑ — Asian↓	.65**	.18	.10	.16
9. ePortfolio: Holistic Score (White = 89) (Asian = 59)	.33	.13	.37*	.74**	.81**	.76**	.68**	.73**	White↑ — Asian↓	.36**	.18	.30**
10. Writing Course Grade (White = 856) (Asian = 498)	.41**	.19**	.20**	.06	.03	.07	.01	.03	.12	White↑ — Asian↓	.43**	.75**
11. Next Writing Course Grade (White = 810) (Asian = 517)	.33**	.26**	.26**	.39**	.41**	.30*	.29*	.14	.30*	.30*	White↑ — Asian↓	.43**
12. Next Semester GPA (White = 937) (Asian = 620)	.46**	.29**	.32**	.18	.16	.15	.09	.11	.18	.66**	.45**	White↑ ←Asian

*$p < .05$
**$p < .01$

Table 7: ***Correlation of Measures: Hispanic and Black Students***

Measures	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
1. HS Rank (Hispanic = 344) (Black = 154)	Hispanic→ Black↓	.09	.21**	.18	.07	.33	.31	.42*	.39	.28**	.39**	.30**
2. SAT Critical Reading (Hispanic = 510) (Black = 243)	.17*	Hispanic↑ — Black↓	.72**	.13	.10	.04	.20	-.04	.20	.11*	.23**	.19**
3. SAT Writing (Hispanic = 510) (Black = 243)	.25*	.67**	Hispanic↑ — Black↓	.11	.01	.00	.23	-.12	.08	.22**	.29**	.28**
4. EPortfolio: Rhetorical Knowledge (Hispanic = 36) (Black = *qns*)	*qns*	*qns*	*qns*	Hispanic↑ — Black↓	.83**	.61**	.57**	.61**	.82**	-.01	.26	.19
5. EPortfolio: Critical Thinking (Hispanic = 36) (Black = *qns*)	*qns*	*qns*	*qns*	*qns*	Hispanic↑ — Black↓	.59**	.53**	.64**	.80**	.10	.45*	.34
6. EPortfolio: Writing Processes (Hispanic = 36) (Black = *qns*)	*qns*	*qns*	*qns*	*qns*	*qns*	Hispanic↑ — Black↓	.58**	.58**	.70*	-.13	.39*	.32
7. EPortfolio: Knowledge of Conventions (Hispanic = 36) (Black = *qns*)	*qns*	*qns*	*qns*	*qns*	*qns*	*qns*	Hispanic↑ — Black↓	.43**	.62**	.08	.42*	.48**
8. EPortfolio: Composing in Electronic Environments (Hispanic = 36) (Black = *qns*)	*qns*	*qns*	*qns*	*qns*	*qns*	*qns*	*qns*	Hispanic↑ — Black↓	.72**	.02	.41*	.49**
9. EPortfolio: Holistic Score (Hispanic = 36) (Black = *qns*)	*qns*	*qns*	*qns*	*qns*	*qns*	*qns*	*qns*	*qns*	Hispanic↑ — Black↓	.08	.31	.41*
10. Writing Course Grade (Hispanic = 391) (Black = 199)	.37**	.06	.13	*qns*	*qns*	*qns*	*qns*	*qns*	*qns*	Hispanic↑ — Black↓	.41**	.66**
11. Next Writing Course Grade (Hispanic = 403) (Black = 201)	.20*	.23**	.27**	-.30	-.22	.17	-.22	-.20	-.05	.32**	Hispanic↑ — Black↓	.44**
11. Next Semester GPA (Hispanic = 485) (Black = 243)	.43**	.27**	.27**	*qns*	*qns*	*qns*	*qns*	*qns*	*qns*	.64**	.40**	Hispanic↑ ↑ ←Black

*$p < .05$
**$p < .01$
Note: Sample sizes under 30, too small for inferential analysis, are designated *qns* (quantity not sufficient)

In both the overall sample shown in Table 1 and all sub-groups shown in Tables 4 through 7, ePortfolio scores reached medium-to-high statistically significant correlations with each other. This important finding provides perspective of the relationship of ePortfolio scores to pre-college measures, enrolled college measures, and predictive college measures.

Regarding criterion evidence related to pre-college measures in the overall sample, high school rank demonstrates low statistically significant correlations with all ePortfolio scores except critical thinking. With the exception of female students whose ePortfolio scores correlate at medium-to-high statistically significant levels with high school rank, statistically significant correlations are largely absent in white, Asian, and Hispanic students. In terms of SAT Critical Reading relationships with ePortfolio scores, statistically significant correlations remain low in the overall sample, and only a single trait (knowledge of conventions) achieves statistical significance for men in analysis of gender. Correlations are absent in White, Asian, and Hispanic students. In terms of SAT Writing, these patterns of low correlation and absence of statistical significance continue with ePortfolio scores in the overall sample, with no statistically significant relationships present for Hispanic students.

Regarding criterion evidence related to the concurrent academic measure of course grades in the overall sample, low statistically significant correlations are present with ePortfolio scores in critical thinking, writing process, knowledge of conventions, and holistic score. For white students, low-to-medium statistically significant correlations are present for all scores except composing in electronic environments. For male students, low correlations are recorded only for writing processes and knowledge of conventions. No statistically significant relationships are present for female, Asian, or Hispanic students.

Regarding criterion evidence related to next writing course grade in the overall sample, presence of statistically significant correlations is somewhat improved from course grades. With the exception of composing in electronic environments, ePortfolio scores are correlated at low levels with next course grade. For male students, all ePortfolio scores correlate at low levels with next writing course grade. For female students, no positive statistically significant correlations are recorded. For White students, statistically significant correlations are low or absent. For Asian students, the presence of correlations are stronger, with all ePortfolio scores correlating with next course grade at low-to-medium levels, with the exception of composing in electronic environments. Similar patterns of medium statistically significant correlations are present with Hispanic students, with the exception of rhetorical knowledge and holistic score.

Regarding criterion evidence related to next semester GPA in the overall sample, low correlations are present with all ePortfolio scores. This pattern continues with male students. No statistically significant correlations appear for female students between next semester GPA and ePortfolio scores. For White students, low statistically significant correlations are present only for writing processes and holistic score. No statistically significant correlations are present for Asian students between ePortfolio scores and subsequent course grades. For Hispanic students, medium statistically significant correlations are present for knowledge of conventions, composting in electronic environments, and holistic score.

6.3. Fairness: Group Difference and Differential Validity

As prelude to an analysis of fairness, it is important to recall the gender and race/ethnicity differences shown in Table 2. As demonstrated, there are no statistically significant differences on ePortfolio scores of White, Asian, and Hispanic students. The only criterion measure that yields no statistically significant differences among race/ethnicity groups is high school rank. Absence of statistically significant differences among race/ethnicity groups is an important indicator of fairness and provides an important background to analysis concerning differential validity.

Table 8 provides regression models on pre-college, enrolled, and predictive measures. Models are coded according to measures in Table 1. So, for example, Model 1A includes measures 1 (HS Rank) +2 (SAT Critical Reading) +3 (SAT Writing) → 9 (Writing Course Grade). While predictive patterns are shown for pre-college measures to facilitate comparison (Models 1A through 1D), four analyses are of special significance to the present study: ePortfolio trait scores as predictive of holistic score (Model 2A); ePortfolio trait and holistic scores as predictive of course grade (Model 2B); ePortfolio trait and holistic scores as predictors of subsequent course grade (Model 3A); and ePortfolio trait and holistic scores as predictors of next semester GPA (Model 3B).

Table 8: ***Regression Models: Pre-College, Enrolled, and Predictive Measures Predicting ePortfolio Holistic Score, Writing Course Grade, Next Writing Course Grade, and Next Semester GPA***

	All		Male		Female		White		Asian		Hispanic		Black	
	R^2	F	R^2	F	R^2	F	R^2	F	R^2	F	R^2	F	R^2	F
Pre-College Measures														
Model 1A 1+2+3→9	.06*nss*	(3, 108) = 2.11	.03*nss*	(3, 95) = 1.03		*qns*	.08*nss*	(3, 47) = 1.09		*qns*	.065*nss*	(3, 47) = 1.09		*qns*
Model 1B 1+2+3→10	.14*** (.99)	(3, 1076) = 58.73	.12*** (1.03)	(3, 880) = 39.79	.25*** (.82)	(3, 192) = 21.31	.16*** (.99)	(3, 403) = 24.60	.16*** (.86)	(3, 215) = 14.05	.11*** (.99)	(3, 243) = 10.38	.16*** (1.10)	(3, 106) = 6.78
Model 1C 1+2+3→11	.14** (.91)	(3, 1083) = 60.39	.13*** (.93)	(3, 877) = 44.09	.16*** (.81)	(3, 202) = 13.04	.17*** (.89)	(3, 387) = 25.57	.15*** (.79)	(3, 228) = 13.41	.13*** (.90)	(3, 259) = 12.36	.13*** (1.15)	(3, 112) = 5.53
Model 1D 1+2+3→12	.22*** (.77)	(3, 1301) = 119.74	.19*** (.79)	(3, 1058) = 81.68	.35*** (.70)	(3, 239) = 42.80	.24*** (.77)	(3, 448) = 46.90	.24*** (.76)	(3, 282) = 29.30	.15*** (.75)	(3, 315) = 17.77	.27*** (.81)	(3, 137) = 17.17
Enrolled College Measures														*qns*
Model 2A 4+5+6+7+8→9	.83*** (.91)	(5, 204) = 199.16	.84*** (.91)	(5, 173) = 179.59	.77*** (.94)	(5, 25) = 16.34	.82*** (.85)	(5, 83) = 77.73	.82*** (.86)	(5, 53) = 47.90	.80*** (1.09)	(5, 30) = 24.20		*qns*
Model 2B 4+5+6+7+8+9→10	.06* (.84)	(6, 200) = 2.13	.08* (.78)	(6, 170) = 2.44	.09*nss*	(6, 23) = .37	.19** (.81)	(6, 81) = 3.20	.04 *nss*	(6, 51) = .33	.12 *nss*	(6, 28) = .61		*qns*
Predictive College Measures														
Model 3A 4+5+6+7+8+9→11	.09** (.91)	(6, 180) = 2.82	.09* (.94)	(6, 154) = 2.44		*qns*	.11 *nss*	(6, 73) = 1.46	.22 *nss*	(6, 46) = 2.19		*qns*		*qns*
Model 3B 4+5+6+7+8+9→12	.07* (.69)	(6, 196) = 2.38	.08* (.64)	(6, 167) = 2.34		*qns*	.14 *nss*	(6, 78) = 1.94	.05 *nss*	(6, 52) = .42		*qns*		*qns*

*$p < .05$
**$p < .01$
***$p < .001$
Note: p-values not statistically significant at the 0.05 level are designated as *nss*. For statistically significant measures, standard errors are reported in parenthesis under R^2.
Sample sizes under 30 are designated as *qns*. Model designations follow Tables 1 and 2: 1 = HS Rank; 2 = SAT Critical Reading; 3 = SAT Writing; 4 = ePortfolio Rhetorical Knowledge; 5 = ePortfolio Critical Thinking; 6 = ePortfolio Writing Processes; 7 = ePortfolio Knowledge of Conventions; 8 = ePortfolio Composing in Electronic Environments; 9 = ePortfolio Holistic Score; 10 = Writing Course Grade; 11 = Next Writing Course Grade; and 12 = Next Semester GPA.

Consistent with evidence in Tables 4 through 7 that ePortfolio scores reached medium-to-high statistically significant correlations with each other, Table 8 shows that there is a statistically significant, high coefficient of determination for all gender and race/ethnicity groups. At its lowest, Model 2A accounts for 77% of the variance for female students. However, no such predictive power is found in the ability of ePortfolio trait and holistic scores as predictive of course grade in Model 2B. At its best, 19% of the variance is account for by Model 2A for white students; at its worst, the model fails to meet the test of statistical significance for Asian and Hispanic students. For these two student groups, evidence of differential validity is therefore absent, and no claims for such evidence can be made for these students of the ability of ePortfolio scores to predict concurrent course grades.

Regarding evidence related to the ability of ePortfolio scores to predict subsequent writing course grade, statistically significant coefficients of determination are low for the total population and for male students. Model 3A fails to meet the test of statistical significance for White and Asian students. Evidence of differential validity is again absent, and no claims for such evidence can be made for these students of the ability of ePortfolio scores to predict next writing course grades.

Regarding evidence related to the ability of ePortfolio scores to predict next semester GPA, a similar pattern emerges. Statistically

significant coefficients of determination account for only eight percent of the variance for the total population and for male students, and Model 3B fails to meet the test of statistical significance for White and Asian students. Again, evidence of differential validity is absent, and no claims for such evidence can be made for these students of the ability of ePortfolio scores to predict next semester GPA.

7.0 Discussion

Quantitative analysis of ePortfolio scores raises unique complexities. As these are identified, it becomes increasingly important to design research according to evidential categories of reliability/precision, validity, and fairness. When program assessment was at its zenith in US post-secondary institutions, Messick (1995) argued that no form of assessment could claim exceptionalism from evidential categories intended to support score interpretation and use. His observations about performance assessment, then as now viewed as a vehicle for robust construct representation, are worth remembering:

> The principles of validity apply to all assessments, whether based on tests, questionnaires, behavioral observations, work samples, or whatever. These include performance assessments which, because of their promise of positive consequences for teaching and learning, are becoming increasingly popular as purported instruments of standards-based education reform. Indeed, it is precisely because of these politically salient potential consequences that the validity of performance assessment needs to be systematically addressed, as do other basic measurement issues such as reliability, comparability, and fairness. (Messick, 1995, p. 5)

As the present study demonstrates, the same argument may be usefully made for ePortfolio-based assessments. In the case of writing assessment, special advantages are apparent in terms of using known sources of validity such as the WPA Outcomes Statement. An established way to frame instruction and assessment, the WPA Outcomes Statement is a valuable curricular strategy that can be used in assembling evidence related to reliability, validity, and fairness. In the present study, ePortfolios similar to that shown in Figure 1 form the center of the assessment, and Tables 1 through 8 illustrate that valuable evidence may be gathered regarding student learning when instruction and assessment are conceptualized as a single phenomenon. In program assessment, it is not top scores or high correlations that matter as much as the ability to engage in fine-grained analysis of student work that, in turn, leads to identification of additional opportunities to learn—the politically salient potential consequences about which Messick wrote two decades ago.

As we turn to discussion of the study findings presented in terms of the three research questions, it may be useful to consider the foundational measurement categories used in the present study function as threshold concepts—categories of key evidential forms that allow what Yancey (2015) has termed a "portal for planning" (p. xix). Rather than view study findings as interpretation and use arguments aimed towards justification, the six-year instructional and assessment history records what occurs when instruction and assessment work together to structure principled analysis. We find that flexibility, comparison, and purpose are usefully associated with reliability, validity, and fairness in presenting the following generalization inferences regarding the writing ability of the students assessed in the present study.

7.1. Inter-rater Reliability/Precision: Interpretative Flexibility

Levels of inter-rater consensus and consistency evidence presented in Table 3 reveal that standard gauge reliability guidelines are of little use in interpreting ePortfolio scores. If scores from complex writing assessments are to be interpreted and information from them used, then researchers are best served by calling into question the 0.7 correlation coefficient established by writing tasks associated with standardized testing. These limits are recognized in the *Standards* (AERA, APA, NCME, 2014) and deserve emphasis in their application to the present study:

> Each step toward greater flexibility in the assessment procedures enlarges the scope of the variations allowed in replications of the testing procedure and therefore tends to increase the measurement error. However, some of these sacrifices in reliability/precision may reduce construct irrelevant variance or construct underrepresentation and thereby improve the validity of the intended interpretations of the scores. (p. 36)

While it is generally recognized that all educational assessment is a series of trade-offs between validity and reliability, the special case of consensus and consistency estimates regarding ePortfolio score reliability lends specificity to the discussion. High rates of inter-rater reliability are of little value if the construct representation is, as Kane (2006) has written, a "very narrow slice of the target domain of literacy" (p. 31).

Because studies of ePortfolio scoring results are very recent, levels of consistency such as those proposed by White, Elliot, and Peckham (2015, Table 4.3, p. 123) are an evidence-based, flexible approach to score interpretation. Establishing high, middle, and low levels of reliability strength for specific measures such as Pearson and quadratic weighted kappa correlations allows inferences to be made and qualifications to be articulated about overall and sub-group performance. In reference to Table 3, for example, rhetorical knowledge scores reveal medium levels of non-adjudicated reliability and high levels of adjudicated reliability for the

overall student sample; hence, while raters did have a moderate degree of difficulty scoring the variable, a high degree of reliability underlies rhetorical knowledge scores from 2010 to 2012.

Based on the literature review and the results of the present study, the following ranges of Pearson correlation coefficients may serve as a useful guide for score interpretation and use related to rater-to-rater reliability evidence when six-point scales are used. Each level assumes 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. In terms of weighted Kappa coefficients, the following ranges may serve as guides: Non-adjudicated low = 0.1 to 0.22, medium = 0.23 to 0.45, and high = 0.46 to 1.00; adjudicated low 0.1 to 0.27, adjudicated medium = 0.28 to 0.56, adjudicated high = 0.57 to 1.00.

While these interpretations may prove useful, they may also serve as a guide for score use. In the present study, ePortfolio scores for female students on writing processes should not be used for any purposes aimed at drawing inferences about this sub-group because statistical significance was not met. Operationally, this conclusion means this variable was therefore withdrawn from further analysis, as shown in Table 5. In similar fashion, scores for Hispanic students shown in Table 3 on rhetorical knowledge, knowledge of conventions, composing in electronic environments, and the holistic score itself met the test of statistical significance only after adjudication. The level of difficulty reading these ePortfolios suggests that caution should be expressed in making claims about the writing ability of Hispanic students on these variables and that further study is required to investigate the problem associated with scoring those variables with this particular sub-group of students.

Examples such as these shed new light on the broad, provocative question by Moss (1994, 2004), “Can there be validity without reliability?” The correct answer, as given by Mislevy (2004), is that reliability/precision must not be sold short based on surface familiarity with techniques of standard practice and lack of specificity regarding assessment aim and subsequent score use. In terms of the present study, Moss’s question, reframed under Mislevy’s call for specificity, may be more precisely framed as an integrative opportunity: “How can evidence associated with reliability/precision help us support and qualify the inferences we make about students so stakeholders receive information that will, in turn, form valid views about student performance?”

7.2. Validity: Convergent Evidence and Comparative Analysis

Without the twelve writing construct measures identified in Table 1 and used throughout the study, meaningful discussion of quantitative information in this first-generation study would be difficult. As such, the patterns observed in Tables 4 to 7 are best interpreted in terms of other performance-based writing studies using criterion information at United States institutions. When first generation assessments are brought forward, comparative information provides directions for score inferences associated with validity evidence.

In a study of post-secondary first-year student writing, Klobucar, Elliot, Deess, Rudniy, and Joshi (2013, Table 4) found low-to-moderate statistically significant correlations on traditional print portfolios—scored with consensus and consistency estimates similar to those reported in Table 2—between all trait scores and holistic scores in a two-year study (2009, n = 151; 2010, n = 135). In 2009 (n = 151) low-to-moderate statistically significant correlations were observed between course grade and each of five trait scores (0.3 to 0.4, $p < .01$) and between course grade and the holistic score (0.43, $p < .01$). In 2010 (n = 135) low-to-moderate statistically significant correlations were again observed between course grade on each of five somewhat different trait scores (0.35 to 0.5, $p < .01$) and the holistic score (0.43, $p < .01$). In a related postsecondary study (n = 135), Klobucar, Deane, Elliot, Ramineni, Dees, and Rudniy (2011) established a correlation of 0.6 ($p < .01$) between a print portfolio holistic score and course grades (Table 2, p. 111).

As Tables 4 to 7 reveal, these levels of correlations are not established in the present study, with the exception of female students shown in Table 5. A reason for the difference may be found in the statistically significant lower trait and holistic scores on the ePortfolio. While there was no statistically significant difference between 2010 scores and 2012 scores on rhetorical knowledge, critical thinking, and knowledge of conventions, lower 2012 scores of statistical significance were demonstrated on writing processes ($t(251) = 4.18$, $p < .001$) and the holistic score ($t(240) = 2.44$, $p < .01$). As well, a new variable, composing in electronic environments, demonstrated a score lower ($M = 6.34$, $SD = 2.11$) than any recorded in 2010. While identical constructs were present, different score patterns appeared when print portfolios were replaced by ePortfolios.

Results suggest that statistically significant correlations between pre-admission measures (high school rank and standardized test scores) and ePortfolio scores will be medium (0.69 to 0.30) to low (0.29 to 0.00). Significant correlations between ePortfolio scores and admitted measures (writing course grade, subsequent writing course grade, and next semester GPA) will also be medium to low. However, if ePortfolio scores are used to determine final course grade—that is, if the ePortfolio score is given a point value used in grade calculation—the correlation coefficient will increase. In a study of 1,208 ePortfolio holistic scores of first-year college students, Kelly-Riley, Elliot, and Rudniy (2016) reported a high r value of 0.82 ($p < .01$) between ePortfolio holistic scores and writing course grades (Table 5, p. 105). The contrast in findings in terms of the present study may serve as a useful guide for score interpretation and use related to concurrent evidence.

Three benefits may be identified in establishing convergent evidence. First, when related performance measures are analyzed in their relationship to ePortfolio scores, study-site ecologies, especially information gained from response process studies, may be more fully understood. Pre-admission, criterion, and predictive measures are essential to understanding ePortfolio trait and holistic scores, interpreting their meaning, and justifying their use. Second, although the traits may be identical, even at a single institution the use of digital techniques mediates the construct. As Katz & Elliot (2016) have hypothesized, measured constructs are mediated by the way the assessment designers sample the construct; depending on how the assessment designers envision the digital scene of action, there may be differences in what is being measured. Indeed, broadly conceived, ePortfolios might be helpfully understood as blurred genres that, as Cumming (2013) noted in the case of integrated writing tasks, have both benefits of robust construct representation and risks of invoking forms of writing that are emerging and therefore difficult to score. Third, while robust construct measures such as ePortfolio scores admirably capture the writing construct, construct under-representation is nevertheless present, as suggested by the moderate-to-low correlations between ePortfolio scores and course grades. While many variables of the writing construct may be captured in ePortfolio-based assessments, the assessment occasion is highly restricted to the reading period—a narrow view of the student ability that does not compare to the course-long observations made by an instructor. One plausible hypothesis of the present study is that disjuncture between ePortfolio scores and writing course grades is evidence of an instructor's longitudinal view of the writing construct and the limited view provided by episodic assessment.

7.3. Fairness: Purpose, Group Difference, and Differential Validity

As is the case with evidence related to reliability/precision and validity, patterns of evidence are clear but interpretation is complex. Table 2 illustrates an important indicator of fairness: There are no statistically significant differences on ePortfolio scores of White, Asian, and Hispanic students. It is important to note that this absence of score difference is not present for any of the other criterion measures in terms of sub-group comparison. Based on present findings, we hypothesize that increased representation of the writing construct results in increased fairness for all student groups.

In terms of predictive validity, the interpretation is more complex. In Model 3A shown in Table 8, statistical significance was not reached for White or Asian students, and the statistically significant coefficients of determination (R^2) for the overall group and White students were demonstrably lower than those obtained by using high school rank and standardized tests. While logic would argue that the entire GPA in a technological university may not necessarily be highly related to writing ability, that interpretation is belied by the high, statistically significant correlations in Tables 4 to 7 for the overall group and all sub-groups between writing course grade and next semester GPA. And, when the ability of ePortfolio scores to predict present and next course grades is examined, the significant R^2 coefficients shown in Table 8 are either low or fail to achieve statistical significance. Only Model 2B for white students ($R^2 = 0.19$)—the predictive relationship between ePortfolio scores and writing course grade—approaches the levels reported by Zwick (2013). Hence, while there are no statistically significant differences on ePortfolio scores of White, Asian, and Hispanic students shown in Table 2, it does not necessarily follow that R^2 coefficients will be at medium or high levels.

Based on the literature and the results of the present study, the predictive value between ePortfolio scores and course grades—as well as values between ePortfolio scores, present and next writing course grades, and next semester GPA—will be low and will probably not exceed statistically significant R^2 coefficients of .3. However, as noted above, if ePortfolio scores are used to determine final course grade, the R^2 coefficient will increase (Kelly-Riley, Elliot, & Rudniy, 2016). In fact a notably strong coefficient of determination ($R^2 = 0.67$, (F (1,2015)) = 2415.52, $p < .001$) was demonstrated in that study, one in which the assessment purpose was distinct from the one presented here. These findings serve as a useful guide for score interpretation and use related to predictive evidence.

While the predictive measures involving ePortfolio scores and other criterion measures are difficult to interpret due to low R^2 coefficients, an unwarranted conclusion would be to argue against the use of ePortfolios in program assessment because of their low predictive powers. As Zwick (2013) observed in the case of high school performance, college admissions decisions need not be based on predicted first year grade point average. The aim of admission is not statistical prediction but, rather, student success. By extension, the purpose of the ePortfolios used in the present study is not to obtain statistically significant R^2 coefficients of comparative value to other measures; rather the aim of the ePortfolios is identification of opportunity to learn for students through program assessment.

In fact, it is quantitative assessment pursued in the interest of fairness that led administrators, based on the information reported in this study, to suspend assessment in 2013 and to concentrate on curricular developments, thereby postponing program assessment until 2014. Specifically, the three-year trends raised concerns that the ePortfolios were not performing as well as the print portfolios. While it was difficult to retrieve ePortfolios (students removed them; software settings prevented access beyond a determined date), informal review revealed that the design shown in Figure 1 had been abandoned. Instead of designing according to the WPA Outcomes Statement, students had begun using the ePortfolios as digital filing cabinets with little categorization of work, little experimentation with new digital forms of writing, or little use of reflective statements. Specific concern centered on 2012 trait scores

on writing processes shown in Table 1 in which scores demonstrated a statistically significant decline from 2011 of medium effect size (d = 0.63). Of greater importance were trait scores for composing in electronic environments that similarly demonstrated a decline of medium effect size (d = 0.64). Based on these score patterns and other sources of information, both administrators and instructors were unsure the digital initiatives were represented with equal commitment across sections. Despite the intentions described in 5.4, fears arose that opportunities to learn about writing in digital environments—and the use of ePortfolios to record that effort—were absent. During 2013, additional attention was therefore given to integrating both print and digital forms of writing into the ePortfolios.

Following these efforts, administrators resumed the first-year assessment in the fall of 2014. Scores in 2014 (n = 99) declined at statistically significant levels from 2012 in rhetorical knowledge (M = 7.56, SD = 1.67; $t(215) = 2.84$, $p < .001$; d = .38). No statistically significant difference was observed for scores on critical thinking (M = 7.68, SD = 1.69; $t(215) = 1.50$, p = .13), writing processes (M = 6.71, SD = 1.98; ; $t(203) = -1.19$, p = .23), knowledge of conventions (M = 7.64, SD = 1.52; $t(208) = 1.02$, p = .31), and holistic score (M = 7.4, SD = 2.08; ; $t(209) = .39$, p = .70). Most importantly, scores were raised by statistically significant levels, with an accompanying medium effect size, for composing in electronic environments (M = 7.43, SD = 1.70; $t(215) = -4.15$, $p < .001$; d = .56). Scores suggested that the curricular effort appeared to be on track, and attention at the present writing has turned to continued investigation of score interpretation and use. In comparison to standardized tests used to conduct outcomes-related work similar to our own, it is difficult to imagine ACT's Collegiate Assessment of Academic Proficiency, the Council for Aid to Education's Collegiate Learning, or the Educational Testing Service's Proficiency Profile producing detailed information that would allow curricular investigation of the kind demonstrated in the present study. As Krzykowski and Kinser (2014) have demonstrated, when these standardized tests are in place, we still know little about results, score use, or consequences. In other words, the transparency demonstrated in locally-developed assessments is absent from standardized tests. Perhaps this transparency is what Condon (2013) alluded to in his challenge to large-scale test designers: "Let the low-yield tests match the richer data set these newer tests provide, and let them address the key aspect of test validity that they so often ignore: the consequences of using their tests" (p. 107).

In terms of research design issues and their relationship to fairness, Mislevy (2004) has observed that the challenge to measurement specialists is "to continually broaden their methodology, to extend the toolkit of data-gathering and interpretation methods available to deal with increasingly richer sources of evidence, and more complex arguments for making sense of that evidence" (p. 243). Returning to the use of threshold concepts, we want to close by reminding readers that ePortfolio-based assessment is an activity and a subject of study unified by purpose. Whether used to assess general achievement or proficiency—or to determine specific course alignment, certification of academic ability, formative review of individual student ability, program review, or research—ePortfolio-based assessments vary in purpose. As the comparative evidence from ePortfolios used for certification demonstrates, evidence related to reliability, validity, and fairness will vary. There is no standard gauge, but there are foundations of principled measurement design leading to transparent reporting. The more such principles are followed and refined, the better we will be able to interpret and use scores for the benefit of students. While the challenges are substantial, opportunities to advance knowledge based on foundational measurement principles, with special attention to score disaggregation and reporting transparency, are apparent.

8.0 Lessons Learned

Used for program assessment purposes, our analysis of fundamental issues in ePortfolio assessment raises issues specific to our institution that are, in turn, instrumental of larger post-secondary concerns about longitudinal score interpretation. In terms of research, ePortfolio-based assessments pose new assessment challenges and opportunities. Substantial challenges remain for research involving ePortfolio assessment regarding limits of inferences that can be drawn from reliability, validity, and fairness evidence. Notably, the present study illustrates challenges associated with sample size, design standards, replication issues, measurement of fairness, and reporting transparency. Potential directions for resolution of these challenges are also noted.

8.1. Sample Size

While the sampling plan identified in 5.3 provides a general ability to analyze ePortfolio trait and holistic scores for an overall group, as shown in Table 1, the design is insufficient for performing sub-group analysis. Even over a three-year time span, as Table 3 reveals, an insufficient sample size prevented inferential analysis for scores of Black students. As Table 8 demonstrates, insufficient sample size also prevented predictive analysis for female and Hispanic students. This absence poses a serious limit to the present study, frustrating efforts of the community to learn from the performance of these students whose high school rank and standardized test scores, as Table 8 demonstrates, are the second highest predictor of second semester GPA. A similar problem exists for women students whose small sample size prevented any predictive analysis. In that the ePortfolio scores of women were higher at statistically significant levels than were the scores of male students, as Table 3 shows, further examination of these scores would have benefitted the entire community.

As related to power analysis—the probability that a test will identify a genuine effect—sample size poses a substantial barrier to score interpretation and use. Ellis (2010) has shown that if an effect size of d = .50 is sought and studies were run with 60

participants, statistical significance would be achieved less than half the time. As the 2010 and 2011 ePortfolio scores in Table 1 illustrate, the samples in the present study do not reach that sample size for the entire population, let alone for sub-groups. As Ellis further demonstrates, a sample of 100 would be needed to identify an effect size of $d = .50$ (Table 3.2, p. 64). In the case of the present study, detection of such an effect size would mean that 100 randomly sampled students would have to be identified for each sub-group shown in Table 3 for a total of 600 students. Such a sample size would be well beyond the resources of the host institution and, in reality, would demand prohibitive resource allocation for many post-secondary institutions. Even if the sample size ($n = 64$) specified by Cohen (1992) as necessary for $\alpha = .05$ were to be used—a method not recommended by Ellis—a sample of 384 would still pose a substantial challenge (Table 2, p. 158). Under such power requirements, the effect size reported for ePortfolio trait scores reported in 7.1 are necessarily qualified.

Sampling plan design is a critical consideration to the future of ePortfolio measurement. If power analysis is not possible for annual readings, at least one other alternative is possible. Power analysis could be used to designate a specified confidence interval that could then be accompanied by purposive samples to include at least 30 ePortfolios for each sub-group. For the case at hand, that improved design would have required that 180 ePortfolios be read over a four day period; and, over a three year period, the sample would reach that identified by Ellis. While additional resources would be needed for such a reading, the analytic gains related to power and effect size would be substantial. Indeed, if another day of research could be devoted to analysis of reliability through qualitative techniques, important gains could be made regarding the reasons why some ePortfolios are difficult to score. This extended assessment period would allow use of multi-method design techniques long held by Haswell (1998) to be profitable in assessment validation.

Although shortcomings of the present sampling plan defined in 5.3 are clear, the design nevertheless allows the generalization inferences made in 7.1, 7.2, and 7.3 to be made from ePortfolio scores to the larger sample. While these generalization inferences are limited, they can nevertheless be used to increase the body of knowledge on ePortfolio-based assessment and to allow individual institutions to advance opportunity to learn for all students.

8.2. Design Standards

In the United States, an initiative of the U.S. Department of Education's National Center for Education Evaluation and Regional Assistance produced a procedures and standards handbook aimed at establishing "rigorous and relevant research, evaluation, and statistics" to improve the nation's education system (Institute of Education Sciences, 2014, p. 1). To meet the design standards without reservation, an intervention study must employ a randomized control design, avoid both overall attrition (i.e., the rate of attrition for the entire sample) and differential attrition (i.e., the difference in the rates of attrition for the intervention and comparison groups), establish baseline equivalence for all groups, demonstrate that reliability and validity evidence be collected for both intervention and comparison groups, and account for confounding factors that limit the study (i.e., all of the intervention group classes are from a single institution). Regarding single case studies that do not employ a randomized control design, such as the present study, an independent variable must be systematically manipulated, inter-assessor agreement must be demonstrated on 20% of the data collected, and attempts must be made to demonstrate effect over time for the study to meet the design standards.

As the present study demonstrates, it is difficult, if not impossible, for any institution to meet the intervention design standards; and, while the single case study design standards are addressed in the present study, there is no deliberate manipulation of the ePortfolio traits and holistic score. As Weiss (1998) has noted in the criticism of such designs, the experiment demands that the program remain constant as the predictor variable is manipulated—a situation that constrains instructors, ever desirous of change for the good of their students, to the extent that the meaning of the results is unclear. While the effort of the Institution of Education Sciences may be seen as an extreme case of misplaced rigor, it is important to recognize that such standards reflect known methodological challenges and, in many cases, would prohibit the peer review that is necessary for publication of ePortfolio research and subsequent classroom impact.

8.3. Replication Issues

As Brennan (2001) has noted, the concept of replication is central to an understanding of reliability. Of course, reliability is a property of scores. It is not a property of assessments, nor is it a property of statistics. Since many kinds of reliability evidence are needed to determine score consistency, replication is needed. Once the need for replication is established, we can better understand the standards set by the Institute of Education Sciences (2014) in terms of emphasis on demonstrated effect over time. If our colleagues are indeed misplacing their notion of rigor, that misdirection is likely due to underestimating the complexities of performance assessments of complex constructs such as writing. As Brennan (2001) recognized, "From a methodological perspective, I believe that performance assessments have had at least one very positive impact—they have forced investigators to recognize the role of multiple facets in characterizing measurement procedures" (p. 307). Among those facets are the very issues of reliability, validity, and fairness raised in the present study, especially as they are related to sampling plan design. While item response theory, for example, may more readily address issues of replication with multiple-choice formats, performance assessments do not readily yield to such methods because of the limited number of scores available. As Brennan noted, sample sizes are also a "part of the story" (p. 308). If this story is to have a happier ending in terms of replication, then smaller, more

compartmentalized constructed response tasks might be used. Exemplary here is CBAL™ initiative in its use of classical test theory, generalizability theory, and item response theory to evaluate parallel test forms (van Rijn, Chen, & Yan-Koo, 2016).

In terms of replications using complex assessment methods, Kelly-Riley and Elliot (2016) have demonstrated that emphasis on foundational measurement categories of reliability, fairness, and validity allows ePortfolio techniques to be used across different institutional settings. Analysis of scores can then provide institution-specific evidence that can be used to support student learning. These techniques, alert to campus ecologies, are aligned with the central point made by Brennan (2001) regarding quantitative assessment, one that is worth quoting in full:

> Mathematics and statistics provide powerful tools for examining the syntactical elements of measurement issues, especially reliability, but mathematics and statistics per se cannot answer semantic questions about measurement. In particular, in my opinion, there are no meaningful answers to reliability questions without explicit consideration of the nature of replications (intended and actual) of a measurement procedure. It follows that a coherent framework for conceptualizing, interpreting, and estimating reliability requires answering the question, "What constitutes a replication of a measurement procedure?" In answering this question, there is no more important consideration than specifying what is fixed and what is not. (p. 313)

If the present study and the study reported by Kelly-Riley and Elliot (2016) are understood to be replications, then replication must be understood both in terms of effect over time and planned use of foundational categories of evidence. What is fixed, then, must be these categories. Without them, replication will be of little meaning.

8.4. Measurement of Fairness

While the classical model established by Cleary (1968) and focusing on differential validity informs our study, there are readily identifiable limits to that method in terms of evidence related to fairness. As Table 8 suggests, “a test may be fair in predicting performance, but nevertheless predict performance rather poorly” (Hartigan & Wigdor, 1989, p. 255). An alternative to emphasis on prediction equations is an emphasis on realized performance. As opposed to test fair, discussed in 2.2.3, Hartigan & Wigdor (1989) termed these assessments as performance fair (p. 255). Analyses such as that offered by Peterson and Novick (1976), noting the advantages of the Threshold Utility Model, offer decision systems that account for values in which the use of cut scores are established as the value providing the greatest benefit across groups.

Yet even these models overemphasize test design rather than test use, as Cole and Zieky (2001) suggested. In the present, as they correctly predicted, fairness must address both issues of design and issues of score use. In terms of design, discussion must revolve around the capability of the assessment to allow individual test-takers the opportunity to demonstrate their abilities on the construct at hand. As Willingham & Cole (1997) wrote, “fair test design should provide examinees comparable opportunity, insofar as possible, to demonstrate knowledge and skills they have acquired that are relevant to the purpose of the test” (p. 11). In terms of use, Willingham and Cole (1997) are insightful: “Fair test use should result in comparable treatment of examinees by avoiding adverse impact due to incorrect inferences or inappropriate actions based on scores or other information normally used with scores” (p. 11). Based on a unified vision, Cole and Zieky (2001) raised four issues that constitute what they term the “essential future facts of fairness”: ameliorating group differences, providing opportunity to perform, deterring misuse, and accommodating individual differences. The prescience of this direction can be demonstrated in *Fairness in educational assessment and measurement*, a collection edited by Dorans and Cook (2016).

Recent research has demonstrated the significance of associating writing assessment with fairness and justice (Kelly-Riley & Whithaus, 2016; Poe & Inoue, 2016). Empirically based, these new theoretical conceptualizations extend the vision of Cole and Zieky (2001) to advance fluid conceptualizations of the least advantaged, structure opportunity to learn for increasingly diverse student populations, define the complex nature of consequence, and understand relationships between group and individual differences and writing construct representation. The challenges to adopting fairness as the first virtue of writing assessment are substantial (Elliot, 2016). Gains related to such theory-building are nevertheless to be realized: designing writing assessment episodes in terms of categories of evidence; reconceptualizing standards to prevent their reduction to technical discussions of methodology; and leveraging the beneficial framework of score interpretation and use as related to the advancement of opportunity to learn.

8.5. Reporting Transparency

The present study identifies both the benefits and challenges involved in ePortfolio research as related to program assessment. A major finding of this study is that are no statistically significant differences on ePortfolio scores of White, Asian, and Hispanic students. In terms of construct validity, this finding supports the inference that increased representation of the writing construct may result in increased fairness for all student groups. When we consider the history of standardized testing and its relationship to disparate impact (Kidder & Rosner, 2002), we realize the significance of our finding.

The study also demonstrates that the presence of rich construct representation such as that demonstrated by ePortfolios is, taken by itself, insufficient for meaningful score interpretation and use. When information related to reliability and fairness is also gathered, the narrative becomes complicated—a complexity that is as nuanced as the students themselves. We must nevertheless continue—indeed, desire—to deal with multiple sources of criterion-related evidence and make sense of that evidence based on score disaggregation. In US postsecondary education, researchers must be able to engage with referential frames valued by the Institute of Education Sciences and other federal and regional agencies to shape research that will be needed to advance equity for the rapidly shifting demographic student populations (Hussar & Bailey, 2013).

None of these demands means that the promises of portfolio-based assessment identified by Hamp-Lyons and Condon (1993) are lost in terms of classroom pedagogy. In fact, just the opposite is true. As our study demonstrates, it is clear that students and instructors were having difficulties adapting to digital writing demands. When administrators decided to pause the assessment and focus on pedagogy, that strategic move was an important decision based on planned evidential categories. Without the information reported in Tables 1 through 8, it would have been difficult to identify the need for proficiency in a specific trait—composing in electronic environments—that held the key to student success in the classroom.

While this study emphasizes the measurement aspects of what we hope will become next generation ePortfolio assessments, our focus is intended to illustrate the important role that quantitative analysis must play. The more we are able to gather evidence related to reliability, validity, and fairness, the more we will be able to remain alert to measurement demands, acknowledge limits associated with our research, establish exemplar reporting standards, and posit new arguments for making sense of complex assessments. In finding our way, we will advance opportunity to learn for all students. Although the language used and methods reported here are limited, they provide an important step for new language and innovative techniques that must be created if we are to structure opportunities for diverse groups of students as they engage the complexities of academic and workplace writing.

Author Note

Alex Rudniy is Assistant Professor in Computer Science in the Gildart Haase School of Computer Sciences and Engineering at Farleigh Dickinson University. He is Co-Principal Investigator on NSF Award 1544239, *Collaborative Research: The Role of Instructor and Peer Feedback in Improving the Cognitive, Interpersonal, and Intrapersonal Competencies of Student Writers in STEM Courses*.

Perry Deess is Director of Director of Planning and Accreditation at New Jersey Institute of Technology. With John Gastil, Philip J. Weiser, and Cindy Simmons, he is co-author of *The Jury and Democracy How Jury Deliberation Promotes Civic Engagement and Political Participation* (Oxford University Press, 2010).

Andrew Klobucar is Associate Professor of English in the Department of Humanities at New Jersey Institute of Technology. He is presently completing *The Algorithmic Impulse: Programmable Writing and the Aesthetics of Information* for University of Alabama Press.

Regina Collins is Associate Director for Assessment and Evaluation in the Office of Institutional Effectiveness at New Jersey Institute of Technology. Creator of WebPAA, she has published papers on topics including privacy in social networks, technology-enhanced learning, and the use of social media in education.

Sharla Sava is Senior Associate and Instructional Designer for Agency Field Development and Prospecting at New York Life in Manhattan. Former director of the Communication Studio at New Jersey Institute of Technology, she is a specialist in incorporating instructional design principals to develop curriculum and improve the design and delivery of course content.

Acknowledgements

We thank New Jersey Institute of Technology Provost and Senior Vice President Fadi P. Deek, and NJIT President Joel S. Bloom for their support of teaching and assessing writing over the past quarter century. As part our program of research in Automated Writing Evaluation beginning in 2009, we would also like to thank Brent Bridgeman, Paul Deane, and Chaitanya Ramineni. While AWE research is not part of the present study, our colleagues nevertheless reviewed the present manuscript for its technical quality. At *JWA*, we would especially like to thank Diane Kelly-Riley and Carl Whithaus for their continued support of our research, including recruitment of the two anonymous reviewers who worked hard to make this manuscript better. The study published here should be understood as a companion piece to Kelly-Riley & Elliot (2016).

References

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

Baird, J-A., Hopfenbeck, T. N., Newton, P., Stobart, G., & Steen-Utheim, A. T. (2014). *Assessment and learning: State of the field review.* Oslo, NO: Knowledge Center for Education. Retrieved from http://www.forskningsradet.no/servlet/Satellite?c=Rapport&cid=1253996755700&lang=en&pagename=kunnskapssenter%2FHovedsidemal

Behm, N., Glau, G., Holdstein, D. H., Roen, D., & White, E. M. (2013). *The WPA Outcomes Statement—A decade later*. Anderson, SC: Parlor Press.

Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), *Construction vs. choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1-27). Hillsdale, NJ: Erlbaum.

Berry, C. M. (2015). Differential validity and differential prediction of cognitive ability tests: Understanding test bias in the employment context. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*, 435-465.

Bridgeman, B., Ramineni, C., Deane, P., & Li, C. (2014). *Using external validity criterion as alternate basis for automated scoring.* Paper presented at the meeting of the National Council on Measurement in Education, Philadelphia, PA.

Bryant, L. H., & Chittum, J. R. (2013). Eportfolio effectiveness: A(n ill-fated) search for empirical evidence. *International Journal of ePortfolio, 3*(2), 189-198. Retrieved from http://www.theijep.com/pdf/IJEP108.pdf

Cambridge, D., Cambridge, B., & Yancey, K. (2009). *Electronic portfolios 2.0: Emergent research on implementation and impact*. Serling, VA: Stylus Publishing.

Camili, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221-256). Westport, CT: American Council on Education/Praeger.

Camp, R. (1983, March 17). *The ETS writing portfolio: A new kind of assessment.* Paper presented at the Conference on College Composition and Communication, Detroit, MI. Educational Testing Service Archives, Princeton, NJ.

Chen, H. T. (2015). *Practical program evaluation* (2nd ed.). Los Angeles, CA: Sage Press.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5*(2), 115-124.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Cole, N. S., & Zieky, M. J. (2001). The new faces of fairness. *Journal of Educational Measurement, 38*(4), 369-382.

College Board. (2012). *State profile report: New Jersey*. New York, NY: College Board.

Collins, R., Elliot, N., Klobucar, A., & Deek, F. P. (2013). Web-based portfolio assessment: Validation of an open source platform. *Journal of Interactive Learning Research, 24*, 5-32.

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing, 18*, 100-108.

Condon, W., Iverson, E. R., Manduca, C. A., Rutz, C., & Willett, G. (2015). *Faculty development matters: Connecting faculty learning to student learning*. Bloomington, IN: Indiana University Press.

Conference on College Composition and Communication. (2015). *Principles and practices in electronic portfolios.* Retrieved from

http://www.ncte.org/cccc/resources/positions/electronicportfolios

Council of Writing Program Administrators. (2000/2008). *WPA Outcomes Statement for first-year composition*. Retrieved from http://wpacouncil.org/positions/outcomes.htm

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly, 10*, 1-8.

Dorans, N. J., & Cook, L.L. (Eds.) (2016). *Fairness in educational assessment and measurement*. New York, NY: Routledge.

Dryer, D. B., Bowden, D., Brunk-Chavez, B., Harrington, S., Halbritter, B., & Yancey, K. B. (2014). Revising FYC outcomes for a multimodal, digitally composed world: The WPA Outcomes Statement for First-Year Composition (Version 3.0). *WPA: Writing Program Administration, 38*, 127-41.

Dysthe, O. Engelsen, K. S. & Lima, I. (2007). Variations in portfolio assessment in higher education: Discussion of quality issues based on a Norwegian survey across institutions and disciplines. *Assessing Writing, 12*(2), 129-148.

Elliot, N. (2015). Validation: The pursuit. [Review of *Standards for Educational and Psychological Testing*, by American Educational Research Association, American Psychological Association, and National Council on Measurement in Education]. *College Composition and Communication, 66*(4), 668–685.

Elliot, N., Briller, V., & Joshi, K. (2007). Quantification and community. *Journal of Writing Assessment, 3*, 5-29. Retrieved from http://www.journalofwritingassessment.org/archives/3-1.2.pdf

Elliot, N., Deess, P., Rudniy, A., & and Joshi, K. (2012). Placement of students into first-year writing courses. *Research in the Teaching of English, 46*(3), 285–313.

Ellis, P. D. (2009). *Effect size calculators.* Retrieved from http://www.polyu.edu.hk/mm/effectsizefaqs/calculator/calculator.html

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, Meta-analysis, and the interpretation of research results*. Cambridge, UK: Cambridge University Press.

Greeno, J. G., & Gresalfi, M. S. (2008). Opportunities to learn in practice and identity. In P. A. Moss, D. C. Pullin, J. P. Gee, E. H. Haertel, & L. J. Young (Eds.), *Assessment, equity, and opportunity to learn (pp. 170-199).* Cambridge, UK: Cambridge University Press.

Hamp-Lyons, L. (2016). Purposes of assessment. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 13-27). Berlin, DE: De Gruyter Mouton.

Hamp-Lyons, Liz, & Condon, W. (1993). Questioning assumptions about portfolio-based assessment. *College Composition and Communication, 44*(2), 176-190.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65-110). Westport, CT: American Council on Education/Praeger.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*, 23-34. Retrieved from http://www.tqmp.org/Content/vol08-1/p023/p023.pdf

Harrington, S., Rhodes, K., Fischer, R., & Malenczyk, R. (2005). *The outcomes book: Debate and consensus after the WPA Outcomes Statement*. Logan, UT: Utah State University Press.

Hartigan, J.A. & Wigdor, A.K. (1989). Fairness in employment testing: Validity generalization, minority issues and the General Aptitude Test Battery. Washington, DC: National Academy Press. Retrieved from https://www.nap.edu/catalog/1338/fairness-in-employment-testing-validity-generalization-minority-issues-and-the

Hussar, W. J., & Bailey, T. M. (2013). *Projections of education statistics to 2022* (NCES 2014-051). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office. Retrieved from http://nces.ed.gov/pubs2014/2014051.pdf

Institute of Education Sciences. (2014). *What Works Clearinghouse™: Procedures and standards handbook version 3.0*. Washington, DC: U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/Handbooks

Isaacs, E., & Knight, M. (2013). Assessing the impact of the Outcomes Statement. In N. Behm, G. Glau, D. H. Holdstein, D. Roen, & E. M. White (Eds.), *The WPA Outcomes Statement—A decade later* (pp. 285-303). Anderson, SC: Parlor Press.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education/Praeger.

Kane, M. T. (2013). Validating the interpretation and uses of test scores. *Journal of Educational Measurement, 50*, 1-73.

Kane, M. T. (2015): Explicating validity. *Assessment in Education: Principles, Policy & Practice, 23*(2), 1-14.

Katz, I. R., & Elliot, N. (2016). Information literacy in digital environments: Construct mediation, construct modeling, and validation processes. In B. D'Angelo, S. Jamieson, B. Maid, & J. R. Walker (Eds.), *Information literacy: Research and collaboration across disciplines* (pp. 97-116). Boulder: University Press of Colorado. Retrieved from http://wac.colostate.edu/books/infolit/chapter5.pdf

Kelly-Riley, D., & Elliot, N. (2014). The WPA Outcomes Statement, validation, and the pursuit of localism. *Assessing Writing, 21*, 89-103.

Kelly-Riley, D. Elliot, N. and Rudniy, A. (2016). An empirical framework for ePortfolio assessment. *International Journal of ePortfolio, 6*(2), 95-116. Retrieved from http://www.theijep.com/pdf/IJEP224.pdf

Kidder, W. C., & Rosner, J. (2002). How the SAT creates built-in-headwinds: An educational and legal analysis of disparate impact. *Santa Clara Law Review, 43*, 131–211.

Klobucar, A., Deane, P., Elliot, N., Ramineni, C., Deess, P., & Rudniy, A. (2012). Automated essay scoring and the search for valid writing assessment. In C. Bazerman, C. Dean, J, Early, K. Lunsford, S. Null, P. Rogers, & A. Stansell (Eds.), *International advances in writing research: Cultures, places, measures* (pp. 102-119). Fort Collins, Colorado: WAC Clearinghouse/Anderson, SC: Parlor Press. 103-119. Retrieved from http://wac.colostate.edu/books/wrab2011/chapter6.pdf

Klobucar, A., Elliot, N., Deess, P. Rudniy, O., & Joshi, K. (2013). Automated scoring in context: Rapid assessment for placed students. *Assessing Writing, 18*, 62–84.

Krzykowski, L. & Kinser, K. (2014) Transparency in student learning assessment: Can accreditation standards make a difference? *Change: The Magazine of Higher Learning, 46*(3), 67-73.

Lam, R. (in press). Taking stock of portfolio assessment scholarship: From research to practice. *Assessing Writing*. Retrieved from http://www.sciencedirect.com/science/article/pii/S1075293516300514

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387-431). Westport, CT: American Council on Education/Praeger.

Lutkus, A. D. (1985). Testing and reporting on graduates: The New Jersey Basic Skills Assessment Program. *New Directions for Teaching and Learning, 24*, 7-15.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York, NY: Routledge.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8.

Middle States Commission on Higher Education (MSCHE). (2015). *Proposed changes to MSCHE accreditation processes and accreditation cycle*. Philadelphia, PA: MSCHE.

Milewski, G, B., Johnsen, D., Glazer, N., & Kubota, M. (2005). *A survey to evaluate the alignment of the new SAT Writing and Critical Reading sections to curricula and instructional practice*. RR 2005-1. New York, NY: College Board. Retrieved from https://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-2005-1-evaluate-alignment-new-sat-writing-

critical-reading-curricula.pdf

Mislevy, R. J., (2004). Can there be reliability without "realibility?" *Journal of Educational and Behavioral Statistics, 29*(2), 241-244.

Moss, P. (1994). Can there be validity without reliability? *Educational Researcher*, *23*(2), 5-12.

Moss, P. A. (2004). The meaning and consequences of "reliability." *Journal of Educational and Behavioral Statistics, 29*(2), 245-249.

Moss, P. A., Pullin, D. C., Gee, J. P., Haertel, E. H., & Young, L. J. (Eds.). (2008). *Assessment, equity, and opportunity to learn.* Cambridge, UK: Cambridge University Press.

National Research Council. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Committee on Defining Deeper Learning and 21st Century Skills, J.W. Pellegrino & M.L. Hilton (Eds.). Board on Testing and Assessment and Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited. *Educational Research Review, 9*, 129-144.

Parsons, M. D. (1997). *Power and politics: Federal higher education policymaking in the*

1990s. Albany, NY: State University of New York Press.

Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models for culture fair selection. *Journal of Educational Measurement, 13*, 3-29.

Poe, M., & Inoue, A. B. (2016). Writing assessment as social justice [Special issue]. *College English*, *79*.

Pullen, D. C., & Haertel, E. H. (2008). Assessment through the lens of "opportunity to learn." In P. A. Moss, D. C. Pullin, J. P. Gee, E. H. Haertel, & L. J. Young (Eds.), *Assessment, equity, and opportunity to learn (pp. 17-41).* Cambridge, UK: Cambridge University Press.

Rhodes, T., Chen, H. L., Watson, C. E., & Garrison, W. (2014). Editorial: A call for more rigorous ePortfolio Research. *International Journal of ePortfolio, 4*, 1-5. Retrieved from http://www.theijep.com/pdf/ijep144.pdf

Roscoe, J. T. (1969). *Fundamental research statistics for the behavioral sciences*. New York, NY: Holt, Rinehart and Winston.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*(4), 1-11. Retrieved from http://PAREonline.net/getvn.asp?v=9&n=4

Suskie, L. (2014). *Five dimensions of quality: A common sense guide to accreditation and accountability*. San Francisco, CA: Jossey-Bass.

van Rijn, P., Chen, J., & Yan-Koo, Y. (2016). Statistical results from the 2013 CBAL English Language Arts multistate study: Parallel forms for policy recommendation writing. *ETS Research Report* (RM-16-01). Princeton, NJ: Educational Testing Service.

Weiss, C. H. (1998). *Evaluation* (2nd ed.). Upper Saddle River, NJ: Prentice Hall

White, E. M. (2001). The opening of the modern era of writing assessment: A narrative. *College English, 63*(3), 306–320.

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

White, E. M., Elliot, N., & Peckham, I. (2015). *Very like a whale: The assessment of writing programs*. Logan, UT: Utah State University Press.

Williamson, D. D., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practices, 31*, 2–13.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Witte, S. P., & Faigley, L. (1983). *Evaluating college writing programs*. Carbondale, IL: Southern Illinois University Press.

Yancey, K. B. (2015). Introduction: Coming to terms: Composition/rhetoric, threshold concepts, and a disciplinary core. In L. Adler-Kassner & E. Wardle (Eds.), *Naming what we know: Threshold concepts of writing studies* (pp. xvii-xxxi). Logan, UT: Utah State University Press.

Yancey, K. B., McElroy, S. J., & Powers, E. (2013). Composing, networks, and electronic portfolios: Notes toward a theory of assessing ePortfolios. In H. A. McKee & D. N. DeVoss (Eds.), *Digital writing assessment & evaluation*. Logan, UT: Computers and Composition Digital Press/Utah State University Press. Retrieved from http://ccdigitalpress.org/dwae

Yancey, K. B., Robertson, L., & Taczak, K. (2014). *Writing across contexts: Transfer, composition, and sites of writing*. Logan, UT: Utah State University Press.

Zwick, R. (2013). Disentangling the role of high school grades, SAT® scores, and SES in predicting college achievement. *ETS Research Report* (RR-13-09). Princeton: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/RR-13-09.pdf