International Journal of ePortfolio 2016, Volume 6, Number 2, 95-116
http://www.theijep.com ISSN 2157-622X

An Empirical Framework for ePortfolio Assessment

Diane Kelly-Riley
University of Idaho

Norbert Elliot
New Jersey Institute of Technology

Alex Rudniy
Fairleigh Dickinson University

This research focuses on ePortfolio assessment strategies that yield important accountability and reporting information. Under foundational categories of reliability, validity, and fairness, we present methods of gathering evidence from ePortfolio scores and their relationship to demographic information (gender, race/ethnicity, and socio-economic status) and criterion variables (admission tests and course grades) as a means for stakeholders to ensure that all students, especially traditionally underserved students, strengthen their connection to the academy. Data is drawn from two sources: University of Idaho first-year writing program's ePortfolio student certification assessment (n = 1208) and its relationship to the State of Idaho's K-20 longitudinal data collection system; and New Jersey Institute of Technology's longitudinal ePortfolio-based first-year writing program assessment (n = 210). Following results and discussion of these two case studies, we conclude by offering guidelines for quantitative reporting based on fairness as a framework for integrative and principled action.

In response to the US's standardized testing movement during the late 1980s and rebooted by the Spellings Commission report in 2006, portfolio assessment helped usher in the powerful capability to combine student learning, faculty evaluation, and documentation of program outcomes. In contemporary higher education landscapes, ePortfolio-based assessments—combining the print tradition of multiple samples of student performance with digital affordances of new genres—have become commonplace. Locally developed and administered, ePortfolios are viewed as congruent with curricular aims at specific institutional sites and are lauded as preferable alternatives to standardized assessments. As Suskie (2009) noted, these construct-rich assessments "can be used in virtually any learning experience" (p. 204) to document both individual student accomplishments and specific course goals across a curriculum. These two uses of ePortfolio based assessment—student certification and program assessment—are the subject of this study.

Rhodes (2011) asserted that ePortfolios, "a powerful, iterative mode for capturing student work and enabling faculty to assess student learning" (para. 3), allow postsecondary institutions to leverage a vast amount of data regarding student learning; consequently, such assessment allows institutions to respond to multiple levels of mandates. First, Rhodes (2011) noted ePortfolio assessment provides a broad means for institutions to respond to the current high-stakes legislative accountability climate focused on measuring student learning. Second, ePortfolios are specifically responsive to shifting accreditation demands of regional or professional organizations: They yield a collection of identifiable student learning artifacts showing that student learning is aligned with faculty demands, and that coursework prepares students for workplace demands.

We agree that portfolio assessment holds a great deal of potential to respond to these general promises and precise claims. The assessment of ePortfolios nevertheless tends to evade educational measurement scrutiny. Despite their widespread use, a dearth of empirically-based inquiry into ePortfolio assessment continues. In their analysis of 118 peer-reviewed journal articles on ePortfolio research, Bryant and Chittum (2013) found that only 15% of the sample focused on outcomes-based research in which student performance was reported.

Appearing first in 2006, the genre of ePortfolio research is relatively new; as such, the tardy application of empirical assessment techniques in ePortfolio research can be partially traced to three reasons. First, widespread access and use of the high speed Internet that is necessary for ePortfolio use is very recent. According to the Organization for Economic Co-Operation and Development (OECD; 2014), the number of adult Internet users in OECD countries increased very recently from fewer than 60% in 2005 to 80% in 2013, with youths reaching 95% during this period. Accompanying this broad usage is a decrease in unit prices and increase in smart devices with data-intensive applications. Second, the interactive elements accompanying Web 2.0—blogs, social networking, video sharing, and wikis, each important to ePortfolio design—are also relatively recent. When the 2006 *Time Magazine* cover featured "you" as the Person of the Year, the designation was accompanied by praise "for seizing the reins of the global media, for founding and framing the new digital" (Grossman, 2006, p. 41). Functioning in an era of technological advancement and media breathlessness, it is no wonder that traditional descriptive and inferential quantitative techniques appear as forms of scrutiny tangential to the gleaming future to come.

Acknowledgment by the educational measurement community that standard gauge techniques used to judge evidence as fit or failing—and new conceptualizations in psychometrics responsive to advancements in digital technologies and cognitive psychology (Mislevy, 2016)—may be the third reason empirical inquiry into ePortfolio assessment is a recent phenomenon. The *Standards for Educational and Psychological Testing* (2014)—published by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME)—asserted that definitions of traditional measurement procedures have broadened considerably, partially in response to scholarship about the merits of portfolio assessment and the widespread implementation of portfolios in traditional print and digital forms. For example, consensus estimates drawn from a timed, impromptu writing sample and used to estimate inter-reader reliability may be higher than that of an ePortfolio, but the latter constitutes a far richer representation of the writing construct. In the cost-benefit analysis accompanying all educational measurement in the accountability environment described by Rhodes (2011), robust construct representation enriches our ability to make important inferences about students. Because many are interested in investigating how ePortfolios can accommodate the complexity of learning for diverse students, it is first necessary to map empirically the landscape—just as Bryant and Chittum (2013) suggested.

These technological developments and educational measurement evolution suggest an important moment in ePortfolio research, and this study both signals the advent of empirical research in this unique form of performance assessment and suggests directions for research reporting. To these dual ends, we begin this study with a literature review of trends in the assessment of complex writing samples; identify foundational measurement concepts of reliability, validity, and fairness; and propose a unification of these concepts under an opportunity to learn framework. We then turn to two case studies—one conducted at the University of Idaho (UI) and the other at New Jersey Institute of Technology (NJIT). As a basis for discussion, the two case studies are used to demonstrate distinct aims (student certification at UI and program assessment at NJIT) and evidence gathering techniques (both descriptive and inferential) suited to those aims. Following a discussion of findings related to our research questions, we conclude by proposing quantitative reporting guidelines for ePortfolios.

Our perspective in this study is drawn from our experiences in the field of Rhetoric and Composition/Writing Studies (Phelps & Ackerman, 2010). As specialists in writing assessment, our experiences evaluating the complex construct of writing allow us to recognize the difficulty of coming to terms with student performance in both print and digital environments. Because ePortfolios allow robust construct representation and pose unique challenges to our field, our experiences in assessment have led us to conclude that quantitative research is an essential approach that yields important information about student ability, particularly about those who are often overlooked or not counted. Informed by our disciplinary stance, our work reported in this article answers the call of Rhodes, Chen, Watson, and Garrison (2014), who asked, “How do we move beyond perceptions and attitudes to explore how ePortfolios can be used to document evidence of student success and achievement of learning outcomes?” (p. 4). To answer their question of agency, we focus on the unique perspectives empirical techniques afford in capturing the complexity of student learning. While tentative, our answers intend to provide a specific direction, based on advancement of opportunity to learn, for the diverse ePortfolio community.

Literature Review

“At the heart of e-portfolio practice research,” Yancey (2009) wrote,

> is a claim about the significance of evidence-based learning. Whether outcomes are programmatically identified or student-designed, the process of connecting artifacts to outcomes rests on the assumption that the selection of, and reflection on, a body of evidence offers another opportunity to learn and a valid means of assessment. At the same time, research has only recently focused on the process of selection and on what counts as evidence. (p. 31)

To establish a research focus, Yancey, McElroy, and Powers (2013) proposed five directions for assessment of ePortfolios: the role of personalization, coherence, reflection, assessment, and web-sensible design. Calling for a new vocabulary and fresh set of practices, Yancey and her colleagues—all leaders in the field of writing studies—provide important directions for evidence-based investigation. Empirically-based quantitative analysis has a distinct place within these directions. We argue that the newly revised foundational measurement concepts articulated in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) allow us to establish an interconnected vision of score interpretation and use based on fairness, and to move beyond mere statistical applications and the reductionism so often associated with empirical quantitative research (Charney, 1996).

Assessment of Complex Writing Samples

Most information about writing assessment has been gained under the carefully controlled experimental conditions often associated with testing (Elliot, 2005). Historically, this narrow view continued until 1983, when Roberta Camp of the Educational Testing Service proposed that portfolio assessment be based on three aims: to provide a comprehensive measure of writing ability that would allow students to demonstrate a wide range of writing experiences; to formulate common assessment tasks and accompanying standards so that student strengths and weaknesses could be evaluated; and to facilitate the transition from secondary to post-secondary institutions by providing information less subject to distortion than that provided by the current application process. The emphasis on construct representation, task and rubric development, and admission and progression use endures, and the academic community took up the challenge to accommodate more complexity in the assessment of writing. From early work at the State University of New York at Stony Brook (Elbow, 1986) to the current program at Washington State University (Kelly-Riley, 2012), portfolio assessment has continued to emphasize connections between instruction and evaluation. The importance of such connections is also widely documented across disciplines and academic programs (Suskie, 2009).

Robust construct representation—accompanied by a need for consideration of assessment consequence—is especially important to writing studies (Behizadeh & Engelhard, 2015). Viewed as a social cognitive construct, writing is a "technology designed to communicate among people" (Bazerman, 2015, p. 11). Writing instruction, and hence writing assessment, is best executed by attending to four domains: cognitive (e.g., genre, task, audience, writing process, problem solving, information literacy, conventions, metacognition), interpersonal (e.g., collaboration, social networking, leadership, diversity, ethics), intrapersonal (e.g., openness, conscientiousness, extraversion, agreeableness, and stability), and physiologic (e.g., nerve, attention, and vision capacity; White, Elliot, & Peckham, 2015). Seen in this way, the empirical assessment research identified by Bryant and Chittum (2013) as outcomes-oriented and affective in design directs attention to issues in construct representation that appear to be similar across disciplinary communities.

Reliability, Validity, and Fairness

While four domains are designed to facilitate representation of the writing construct, three foundational categories of educational measurement—reliability, validity, and fairness—provide methods of obtaining information about those domains. These foundational categories have undergone substantive evolution from their first articulation in the 1966 *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1966), which placed reliability as the most important consideration in test use and separated it from validity, a property of a particular test. The 1999 *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) entirely revised the concepts of reliability and validity, advanced a unified concept of validity as the most important consideration, and situated validity within the use and interpretation of test scores in particular settings. Further, the most recent version of the *Standards* (2014) elevated the concept of fairness to be a foundational consideration for tests, parallel in importance to validity and reliability.

In the present study, we are particularly influenced by theorists who rearticulated the foundational concepts in the revised AERA, APA, and NCME (2014) standards. Haertel (2006) defined reliability as concerned "with how the scores resulting from a measurement procedure would be expected to vary across replications of that procedure" (p. 65). At the present writing, Generalizability Theory (G theory; Brennan, 2001) provides the best, most nuanced framework for reliability, complete with conceptual and statistical tools for analysis. Regarding validity, Kane (2013) conceptualized it as "the process of evaluating the plausibility of proposed interpretations and uses of test scores" (p. 16). As such, it is not the assessment that is validated; rather, the interpretations and uses of the assessment are validated. In order to achieve clear statements of these uses, Kane (2013) advanced the idea of interpretation and use arguments to support inferences derived from scores. Because we believe that opportunity should be linked to definitions of fairness, as noted above, the orientation towards ethical assessment provided by Suskie (2009) is especially helpful: A fair assessment will use tasks that are equally familiar to all and thus advance opportunity to learn (Kelly-Riley & Whithaus, 2016). The measurement community also supports this common sense orientation as the most recent iteration of the standards (AERA, APA, & NCME, 2014) redefined fairness as

> the validity of test score interpretations for intended use(s) for individuals from all relevant subgroups. A test is fair that minimizes the construct-irrelevant variance associated with individual characteristics and testing contexts that otherwise would compromise the validity of scores for some individuals. (p. 219)

While reliability, validity, and fairness often become silos in practice, using fairness as an integrative principle—as we will demonstrate below—allows an agenda for principled investigation and action.

Opportunity to Learn

The subsequent link between assessment and instruction allows direct attention to consequence if fairness is accepted as an integrative principle of assessment. While traditional identification of intended and unintended consequences remains an important facet of assessment (Messick, 1980), emphasis on opportunity to learn, as Suskie noted (2009), establishes a critical link between instruction and assessment.

A primary aim of assessment, Suskie (2009) observed, is the advancement of opportunity to learn, defined as assurance that each student in a course, program, or college has sufficient opportunity to achieve each established curricular goal. As Pullin (2008) has stressed, emphasis on the opportunity to learn is both a reflection of the learning environment and a concept demanding articulated connections between the assessment and the instructional environment. For the assessment to proceed along the lines of fairness, resonance must be demonstrated among the following: the design of the assessment, the opportunity to learn, and the educative intent to improve and continue that learning. This resonance positions score interpretation and use as a vehicle for examining what Gee (2008) identified as the rights of students in terms of opportunity to learn: universal affordances for action, participation, and learning; assurances of experiential ranges; equal access to relevant technologies; emphasis on both information communication and the communities of practice that manage that information; and emphasis on identity, value, content, and characteristic activities associated with language across academic areas.

Associated with the opportunity to learn is identification of those who are least advantaged by the assessment. There are many reasons that opportunity is denied, and thus the pursuit of fairness calls for disaggregation of assessment scores by sex assignment at birth (gender), race/ethnicity, socioeconomic status (SES), and special program enrollment as we demonstrate in Tables 7 through 11. Depending on the writing task at hand, there are many factors—from genre familiarity to digital proficiency—that could result in student disenfranchisement. Identification of membership along a continuum of groups is not intended to obviate racialization processes; rather examination of group differences reveals a long tradition of empirical study that cannot be resolved by identification of economic status, race/ethnicity, or any singular factor. As we demonstrate in the following two case studies, score disaggregation is a fundamental step in allowing us to learn more about the inferences we can make from ePortfolio scores. Put straightforwardly, data from our two case studies will demonstrate how the category of least advantaged is not fixed and that students may, in fact, shift in and out of that designation.

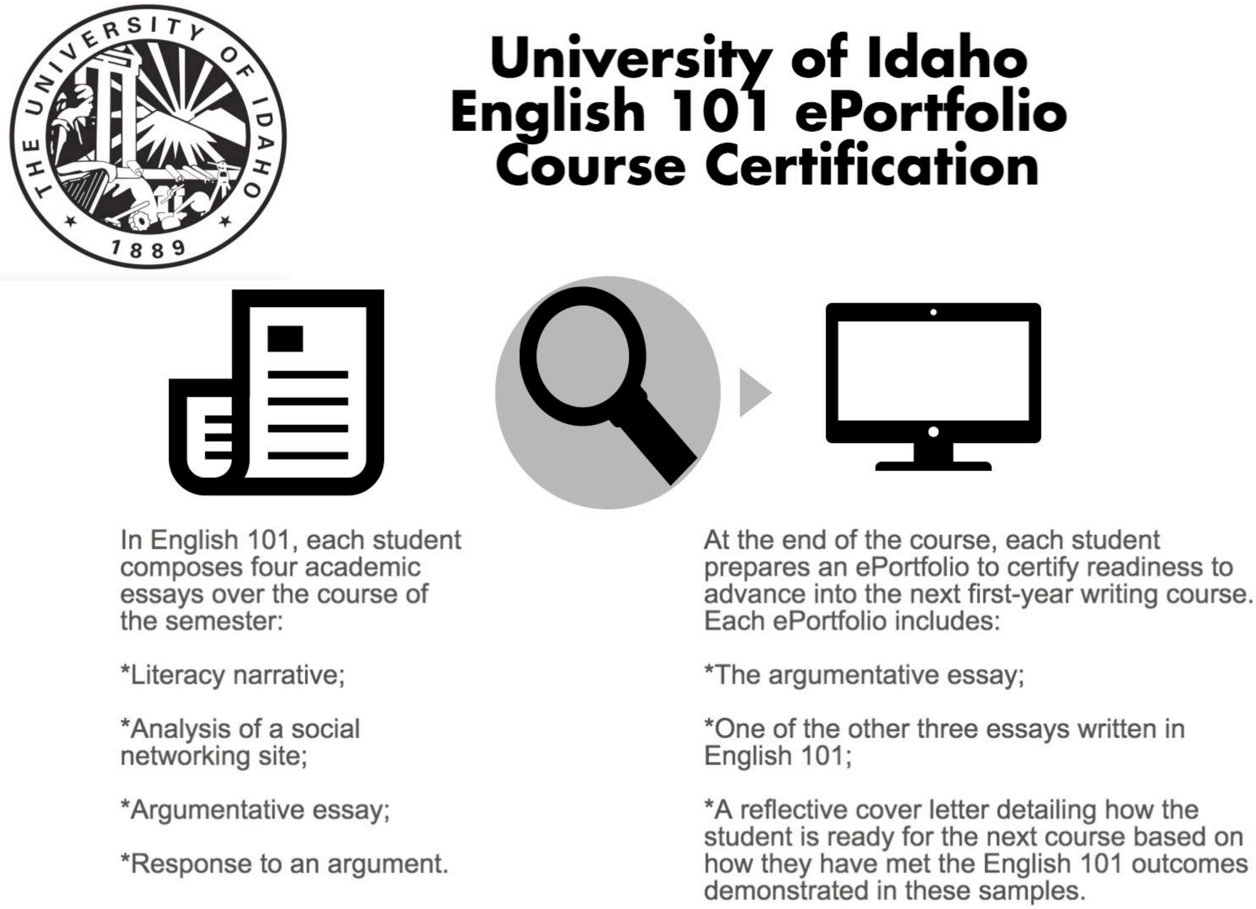
Methods

The two case studies informing the recommendations we make are drawn from distinct intuitions with differing missions. This range demonstrates the universality of the analytic methods we use and the promise of the foundational approach we advocate. We begin with a description of both universities and the ePortfolio-based assessments at each. We then turn to detailed methodological considerations, including identification of criterion measures and sub-group categories, a description of our quantitative techniques, and identification of our research questions.

University of Idaho: ePortfolios and Individual Student Certification

University of Idaho (UI)—the state's land grant, flagship institution—is the one of two study locations. According to the Carnegie Classifications of Higher Education, UI is designated as a Research University with high research activity that selectively admits undergraduate students and has doctoral and professional dominant graduate programs. At University of Idaho, ePortfolios have limited institutional adoption, but the English Department has incorporated end of course portfolio assessment using holistic scores in the first course of the first-year writing sequence, English 101 (Introduction to Academic Writing). The English 101 ePortfolio adapts Belanoff and Elbow's (1986) portfolio assessment as a way for students to certify their readiness to move into the next course in the first-year writing sequence, English 102, College Writing and Rhetoric. ePortfolios have been integral to the UI First-Year Writing Program since 2010 when a standardized curriculum was implemented and administered through the course management system, Blackboard Learn, and the ePortfolio certifies individual student knowledge and skills along with final grades. Each ePortfolio contains the argumentative essay, one of the three other essays written for the course, and a reflective letter that details the student's readiness for English 102 by virtue of meeting the outcomes of English 101 demonstrated in the ePortfolio collection (Figure 1). English 101 is taught by new MA level English graduate teaching assistants, many of whom have little to no teaching

Figure 1
University of Idaho English 101 ePortfolio



University of Idaho
English 101 ePortfolio
Course Certification

In English 101, each student composes four academic essays over the course of the semester:

*Literacy narrative;

*Analysis of a social networking site;

*Argumentative essay;

*Response to an argument.

At the end of the course, each student prepares an ePortfolio to certify readiness to advance into the next first-year writing course. Each ePortfolio includes:

*The argumentative essay;

*One of the other three essays written in English 101;

*A reflective cover letter detailing how the student is ready for the next course based on how they have met the English 101 outcomes demonstrated in these samples.

experience, and the English 101 curriculum is highly structured to mitigate their lack of instructional experience and/or knowledge of writing studies.

At the end of the semester, ePortfolios from English 101 are assessed by teaching assistants, adjuncts, and tenure-line faculty who score student work using the expert reader model of evaluation detailed by Smith (1993) and Haswell and Wyche (1996). Scores are used either to certify students for placement directly into the next course or to decide that the student is not ready for English 102 (and has not passed English 101). Scores from the ePortfolio therefore influence the course grade.

University of Idaho uses the English 101 ePortfolio shown in Figure 1 as a way to gauge student progress through the undergraduate curriculum, and this progress is further analyzed by the state of Idaho's two data systems: Idaho System for Educational Excellence (ISEE), which collects student data in the K-12 setting, and the State Longitudinal Data System (SLDS), which tracks student performance data for all of the postsecondary institutions. Based on data from 2014 to the present, the UI case study highlights the integration of data available through ISEE and SLDS, combined with the ePortfolio assessment at the end of English 101, and reports on the coordination of this data with UI student performance. For the case study reported here, the sample is drawn from 1208 students who enrolled in English 101, Introduction to Academic Writing, in Fall 2014. Of these, 650 were male and 558 were female. In the sample, 860 are White; 153 are Hispanic/Latino (hereafter referred to as Hispanic); 50 were two or more races. African-American, Asian/Pacific Islander, and Native American students comprised the remaining number, but did not have sufficient numbers to conduct the statistical analysis.

NJIT: ePortfolios and Program Assessment

The location of the second study is New Jersey Institute of Technology (NJIT), the state's science and technology institution. Classified as a science, technology, engineering, and mathematics-dominant research institution by the Carnegie Commission on Higher Education, NJIT selectively admits undergraduate and graduate students. Historically, portfolio assessment at NJIT has been used as a form of program assessment—a planned, recurrent documentation effort intended to demonstrate that those responsible for the program have advanced its mission of student learning—in support of accreditation by the Middle States Commission on Higher Education (MSCHE). Featured in the 2002 accreditation process as a print-based evaluation and in 2012 as a digital evaluation, assessment of complex writing samples has supported both successful reaccreditation visits and will be part of the 2017 periodic review report. Evolving from print portfolios (Elliot, Briller, & Joshi, 2007), ePortfolios have been used in first-year writing (Klobucar, Elliot, Dees, Rudiny, & Joshi, 2013), undergraduate technical writing (Johnson & Elliot, 2010), and at the graduate level in professional and technical writing (Coppola & Elliot, 2010). While the Department of Humanities has historically used ePortfolios to benefit NJIT, ePortfolios have had limited institutional adoption, as is the case with UI. In the present study, attention is given to ePortfolio assessment conducted in Humanities 101 (English Composition: Writing, Speaking, Thinking I) and the relationship of those scores to the next writing course (Humanities 102, English Composition: Writing, Speaking, Thinking II).

Conducted annually in the fall with entering first-year students in Humanities 101, the ePortfolio assessment used in this case study is congruent with the proposed MSCHE (2015) annual updates focusing on assessment student learning. ePortfolio assessment is also designed to serve the NJIT's professional programs as accredited by the Association to Advance Collegiate Schools of Business, Accreditation Board for Engineering and Technology, and National Architectural Accrediting Board. In this use, the NJIT ePortfolio system is similar in its aim to that of Ketcheson (2009) and Larkin and Robertson (2013); in both institutions, however, no claim to universal institutional use can be made that is comparable to that of Williams (2010).

In contrast to ePortfolio use at UI, scores from the NJIT ePortfolio shown in Figure 2 do not influence the course grade. To ensure that grade influence does not occur, assessment of ePortfolios occurs after final grades have been posted. Also distinct from the UI program, NJIT first-year portfolios, while required of all students enrolled in the first writing class, are not all read each semester. Based on traditional power analysis techniques designed to yield a specified confidence interval ranging from 0.8 to 0.95 (Kerlinger & Lee, 1999), ePortfolios are read based on both random and purposive sampling techniques designed to allow representation of student groups (White et al., 2015). Use of ePortfolios for regional accreditation and subsequent principles of selection, of course, does not separate the content of the ePortfolio from the very course that supports its creation. Nevertheless, while student certification requires that ePortfolios from each student be read each semester, program assessment does not demand this level of data collection.

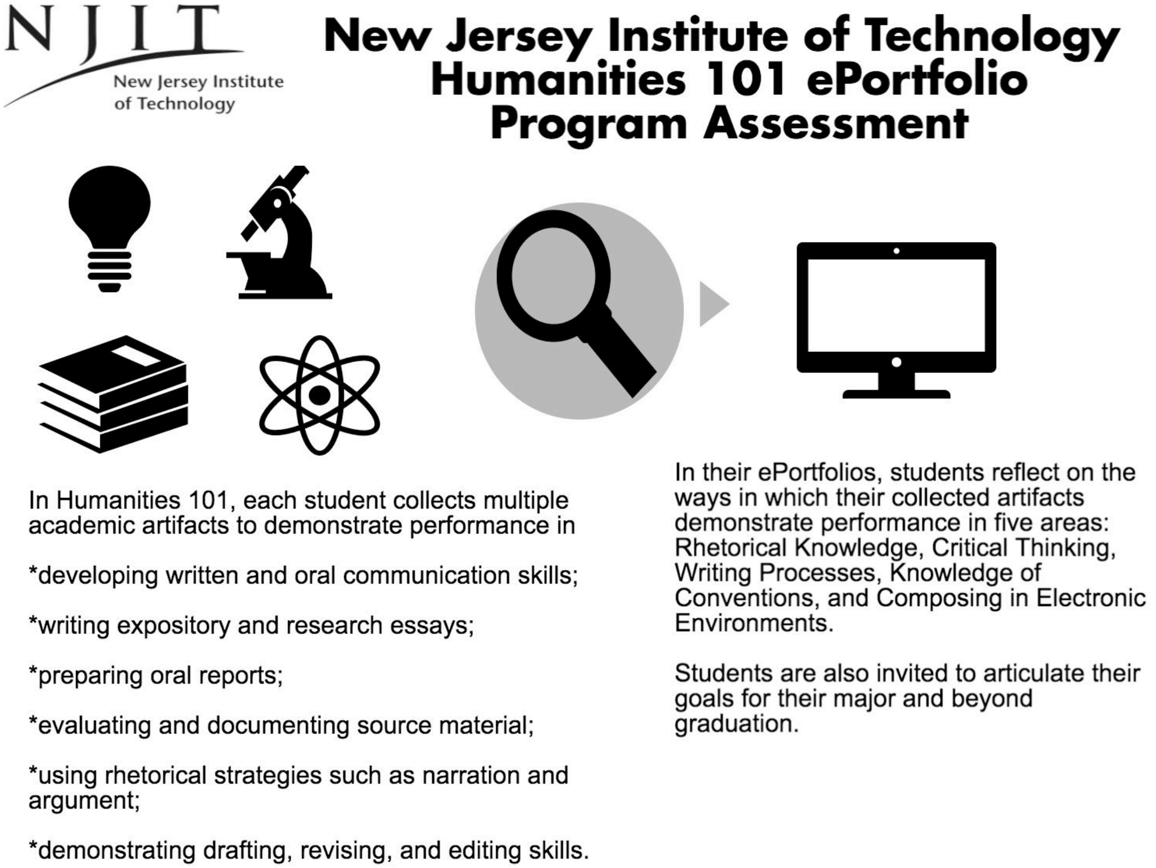
To structure comparison between UI and NJIT, we focus on the NJIT first-year writing sequence from 2010 to the present, with special attention on the formative years for the program from 2010 to 2012. These courses are taught by full-time, experienced lecturers, as well as by tenure-line and tenured faculty, and all who teach the classes participate in the ePortfolio scoring. While the UI ePortfolios are scored holistically, NJIT ePortfolios during this period were scored on a national consensus model of the writing construct (Council of Writing Program Administrators, National Council of Teachers of English, & National Writing Project, 2011). Specifically, trait scores—often termed multiple trait scoring (Hamp-Lyons, 2016)—are provided for rhetorical knowledge, critical thinking, writing processes, and knowledge of conventions (Figure 2). A holistic score is also provided. Differences of scoring method are appropriate to assessment aim. The sole use of holistic scoring at UI is appropriate for the certification purpose of the assessment program. As well, because each ePortfolio must be read, it would be costly and difficult to score each student performance using multiple traits. Because a sample of ePortfolios is read at NJIT, the trait method is appropriate to an aim of identifying student strengths and tailoring the program to leverage success.

Detailed Methodological Considerations

Criterion measures. As an identifiable construct, writing can be measured in a number of ways. Independent of the measure of interest—in this study, ePortfolios—criterion measures are used to evaluate relationships between and among different ways that the construct is represented. Tables 1 and 2 identify pre-college, enrolled, and predictive academic measures important to each university. These include high school grade point average, high school rank, and common standardized test scores used in admissions.

Sub-group categories. Evidence related to reliability, validity, and fairness must be collected for both the overall group and sub-groups relevant to institutions. Tables 1 and 2 identify sub-groups important to the intuitions and relevant to interpretative ranges. While sub-group representation may be small and need qualification in terms of score interpretation and use, it is nevertheless important to collect

Figure 2
NJIT Humanities 101 ePortfolio



New Jersey Institute of Technology
Humanities 101 ePortfolio
Program Assessment

In Humanities 101, each student collects multiple academic artifacts to demonstrate performance in

*developing written and oral communication skills;

*writing expository and research essays;

*preparing oral reports;

*evaluating and documenting source material;

*using rhetorical strategies such as narration and argument;

*demonstrating drafting, revising, and editing skills.

In their ePortfolios, students reflect on the ways in which their collected artifacts demonstrate performance in five areas: Rhetorical Knowledge, Critical Thinking, Writing Processes, Knowledge of Conventions, and Composing in Electronic Environments.

Students are also invited to articulate their goals for their major and beyond graduation.

information on distinct and related group categories. As is clear in the case of NJIT female students, whose ePortfolio sample size was small (n = 31), restricted ranges of this high performing group impact both consistency and correlation evidence.

Idaho is a fairly racially homogenous state, but there is a great deal of economic diversity. Pell grant status is one way to examine students' socio-economic status, but it does not give a range of economic backgrounds. We used the Expected Family Contribution (EFC) as a way to represent a full range of economic information for ePortfolio assessment. Many variables affect students' EFC, and families often are initially referred to consider parental adjusted gross income reported on federal income taxes as a way to estimate anticipated EFC for their college student. Other factors such as assets, requirement account savings, and number of children in college affect EFC. For this project, students' EFC data collected by the University of Idaho was the starting point, and then their EFC was mapped back to a range of adjusted gross income (Onink, 2014); then this adjusted gross income was mapped back to the ten College Board income categories to broadly represent the spectrum of family income (College Board, 2015b, p. 4). Finally, for this analysis, these ten categories were then divided into quartiles. This process ensured that the range of financial background of University of Idaho students was adequately represented, and not simply divided into four equal quartiles. The EFC quartiles divided in the following ways: EFC Quartile 1 = $0-$20,000; EFC Quartile 2 = $20,000-$60,000; EFC Quartile 3 = $60,000-$100,000; and EFC Quartile 4 = $100,000+.

Quantitative techniques. Techniques used in both case studies are descriptive and inferential. Descriptive statistics are used in Tables 1 and 2, and means and standard deviations are shown for all measures. Table 3 uses a Gaussian (normal) distribution to describe consensus scoring techniques. For inferential statistics, general linear modeling is used for the correlation and regression analyses shown in Tables 4 through 11. A confidence level of $p < .05$ is used to ensure that a 95% confidence interval is reached. Interpretatively, the correlation

Table 1
University of Idaho: Descriptive Measures, All Groups

	Pre-College Measures			Enrolled College Measures		Predictive College Measures
	HS GPA (N, M, SD)	SAT Writing (N, M, SD)	ACT Composite (N, M, SD)	ePortfolio: Holistic Score (1-6) (N, M, SD)	Eng 101 Course Grade (N, M, SD)	Eng 102 Course Grade (N, M, SD)
Total	1161, 3.23, .44	919, 457.78, 66.37	594, 20.31, 3.14	1208, 3.73, .95	1208, 2.82, .54	971, 3.21, 1.08
Male	619, 3.13$_{a}$, .44	467, 451. 7$_{a,}$67.45	320, 20.79$_{a}$, 3.27	650, 3.62$_{a}$, 1.01	650, 2.76$_{a}$, .59	501, 3.06, 1.14
Female	542, 3.34$_{a}$, .42	452, 464.03$_{a}$, 64.71	274, 19.76$_{a,}$2.88	558, 3.85$_{a}$, .85,	558, 2.88$_{a}$, .45	470, 3.36, 1.0
White	846, 3.24, .43	678, 465.91$_{a}$, 63.02	449, 20.84, 2.93	860, 3.7, .91	860, 2.84, .51	693, 3.1, 1.1
Hispanic/ Latino	152, 3.12$_{a}$, .45	128, 426.17$_{a}$, 66.05	85, 18.27, 3.22	153, 3.72, .85	153, 2.84, .48	132, 3.1,1.0
Two or More Races	49, 3.36$_{a}$, .42	44, 465.68$_{a}$, 68.79	*qns*	50, 3.64, 1	50, 2.76, .63	43, 3.3, 1.14
First Gen	434, 3.22, .44	354, 447.82, 69	231, 19.97, 3.23	441, 3.80$_{a}$, .894	441, 2.85$_{a}$, .48	360, 3.16, 1.10
Pell Grant	509, 3.18, .44	414, 450, 69.12	275, 19.79, 3.22	516, 3.75, .981	516, 2.80, .55	416, 3.19, 1.07
EFC Q1	313, 3.15$_{a}$, .45	252, 444.60$_{a}$, 68.28	170, 19.55$_{a}$, 3.21	253, 3.74, 1.02	253, 2.79, .56	253, 3.2, 1.06
EFC Q2	253, 3.22, .43	204, 458.53, 69.91	140, 20.34, 3.22	259, 3.78, .93	259, 2.82, .53	202, 3.13, 1.11
EFC Q3	223, 3.27$_{a}$, .43	187, 467.81$_{a}$, 64.6	116, 20.71$_{a}$, 2.93	226, 3.76, .85	226, 2.84, .51	176, 3.3, 1.2
EFC Q4	210, 3.32$_{a}$, .40	179, 472.18$_{a}$, 61.84	110, 21.06$_{a}$, 3.02	211, 3.83, .80	211, 2.88, .46	186, 3.2, 1.07

Note. Different subscripts ($_{a}$) within a row represent *M* different at the 0.05 level by independent sample *t* test (2-tailed) for gender, first generation status, and Pell grant status and by Tukey's HSD test for race and EFCQ. Sample sizes under 30, too small for inferential analysis, are designated *qns* (quantity not sufficient). HS GPA: Male < Female; Hispanic/Latino/a < Two or more races; EFCQ1<EFCQ3; EFCQ1<EFCQ4. SAT writing: Hispanic/Latino/a < Two or more races; Hispanic/Latino/a < White; EFCQ1 < EFCQ3; EFCQ1< EFCQ4. ACT composite: Hispanic/Latino < White; EFCQ1 < EFCQ3; EFCQ1< EFCQ4.

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.

Because we hold that reliability information is an important prerequisite to evidence of validity and fairness, our analysis is presented in terms of reliability, validity, and fairness. However, as we propose, fairness is an important governing concept for both reliability and fairness in advancing the opportunity to learn. Our presentation of information is therefore more functional than conceptual.

Research questions. Our research is guided by the following questions regarding ePortfolio-based assessments used to determine individual and group student performance:

1. How may reliability evidence be used to better understand a general student population and relevant sub-groups in terms of consensus and consistency estimates?
2. How may validity evidence be used to better understand a general student population and relevant sub-groups in terms of correlation analysis?
3. How may fairness evidence be used to better understand a general student population and relevant sub-groups in terms of statistically significant difference and regression analyses?

Results

We begin by describing the first-year writing performance profiles of students at both UI and NJIT. We then proceed to results grouped according to evidential categories of reliability, validity, and fairness. Because of our interest in fairness, additional attention is given to this category. It is important to recall that these are specific types of statistical analyses and are not intended to exhaust the many sources of evidence related to these three foundational measurement concepts. Our report highlights the ways that this framework can be used to examine ePortfolio assessments with different aims: one that certifies student performance and the other for program assessment.

Table 2

NJIT: Descriptive Measures, All Groups

	Total (N, M, SD)	Male (N, M, SD)	Female (N, M, SD)	White (N, M, SD)	Asian (N, M, SD)	Hispanic (N, M, SD)	Black (N, M, SD)
			Pre-College Academic Measures				
HS Rank	1420, 73, 21	1155, $71_{a.}$ 21	265, 80_{a}, 19	502, 72, 21	300, 75, 21	344, 76, 19	154, 72, 19
SAT Writing	2636, 534, 85	2086, 525_{a}, 81	550, 568_{a}, 94	974, 550_{a}, 76	616, 550_{a}, 97	510, 508_{a}, 77	243, 503_{a}, 75
			Enrolled College Measures				
ePortfolio Rhetorical Knowledge	210, 8.06, 2.14	179, 7.94_{s}, 2.22	31, 8.77_{a}, 1.39	89, 8.11, 1.97	59, 8.05, 2.03	36, 8.42, 2.01	*qns*
ePortfolio Critical Thinking	210, 7.88, 2.06	179, 7.73_{a}, 2.09	31, $8.74_{a,}$ 1.57	89, 7.80, 1.94	59, 8.03, 1.9	36, 8.31, 1.93	*qns*
ePortfolio Writing Processes	210, 6.81, 1.96	179, 6.6_{s}, 2.01	31, 7.71_{a}, 1.37	89, 6.62, 1.93	59, 7.15, 1.93	36, 7.06, 1.84	*qns*
ePortfolio Knowledge of Conventions	210, 7.91, 2.02	179, 7.79_{a}, 2.05	31, 8.65_{a}, 1.74	89, 7.96, 1.88	59, 8.05, 1.98	36, 8.22, 1.59	*qns*
ePortfolio Composing in Electronic Environments	210, 6.57, 2.33	179, 6.45, 2.3	31, 7.26, 2.02	89, 6.33, 2.27	59, 6.69, 2.13	36, 6.86, 2.36	*qns*
ePortfolio: Holistic Score	210, 7.6, 2.17	179, 7.46_{a}, 2.21	31, 8.39_{a}, 1.76	89, 7.58, 1.98	59, 7.71, 1.94	36, 8.06, 2.27	*qns*
Hum. 101 Course Grade	2172, 3, 1.09	1727, 2.94_{a}, 1.11	444, 3.24_{a}, .96	856, 3.11_{a}, 1.08	498, 3.1_{a}, 1.0	391, $2.87_{a,}$ 1.06	199, 2.75_{a}, 1.14
			Predictive College Measures				
Hum. 102 Course Grade	2147, 3.11, .96	1678, 3.04, .98	469.34, .93	810, 3.24, .886	517, 3.13, .942	403, 3.01, .967	201, 2.81, 1.11

Note. Different subscripts (a) within a row represent means different by independent sample t test (2-tailed) for gender and by Bonferroni for race/ethnicity. Sample sizes under 30, too small for inferential analysis, are designated *qns* (quantity not sufficient). *p*-values are reported at $p < .05$. Gender: HS rank: M < F; SAT writing score: M < F; ePortfolio rhetorical knowledge: M < F; ePortfolio critical thinking: M < F; ePortfolio writing processes: M < F; ePortfolio knowledge of conventions: M < F; ePortfolio holistic score: M < F; ePortfolio writing course grade: M < F; ePortfolio next writing course grade: Race/ethnicity: H < W, H < A, B < W, B < A; Writing course grade: H < W; H < A; B < W; B < A; Next writing course grade: H < W; B < W; B < A.

Student Profiles

To begin, we highlight results disaggregated by particular demographic characteristics, alternating between findings from the UI and NJIT ePortfolio assessments. Given the extensive amount of data available, we will highlight only key patterns of analysis to illustrate the ways such data can help us understand the complexity of student performance, as viewed through a fairness lens.

Table 1 provides descriptive performance information for various demographic characteristics at UI. To understand how ePortfolios were situated among other measures of student performance, we categorized data as follows: pre-college enrollment measures (high school GPA, SAT Writing scores, and/or ACT composite scores); enrolled college measures (ePortfolio scores and writing course grades); and predictive measures (grades in the next writing course or next semester). The portrait of UI students shown in Table 1 is one that supports the Carnegie Classification description of the university as one of selective undergraduate admission. Compared to state profiles compiled by the College Board (2015a) that include performance on the SAT Writing scores, UI students (n = 919) overall scored above the state sample (*n* = 17, 695, *M* = 442, *SD* = 98) at statistically significant levels (*t*[1162] = 7.26, $p < .001$). In the enrolled and predictive measures, the writing course grade is high in both courses. In the case of ePortfolio holistic scores, the mean score of UI students is above the cut score of 3.0.

As Table 2 demonstrates, NJIT students have profiles similar to those of UI.

Compared to College Board (2015b) state profiles, overall SAT Writing scores of NJIT students from 2010 to 2012 (*n* = 2,636, *M* = 534, *SD* = 85) were higher than state levels (*n* = 85, 012, *M* = 499, *SD* = 118) at statistically

Table 3
University of Idaho ePortfolio Consensus Estimates

	Method: Tier Rating			Efficacy: Score Frequency		
Score level	Tier 1	Tier 2	Final reading	Frequency	%	Cumulative %
Score 6	Distinction	Distinction	Distinction	38	3.1	100
Score 5	Distinction	Pass	Pass	22	1.8	96.9
Score 4	Pass	→	Pass	926	76.7	95
Score 3	No Pass	Pass	Pass	95	7.9	18.4
Score 2	No Pass	No Pass	No Pass	49	4.1	10.5
Score 1	Fail	→	Fail	78	6.5	6.5

Table 4
NJIT ePortfolio Consistency Estimates

	Consistency estimates	
	Method 1: Non-adjudicated Pearson	Method 2: Adjudicated Pearson
ePortfolio: Rhetorical knowledge	.42***	.67***
ePortfolio: Critical thinking	.54***	.71***
ePortfolio: Writing processes	.37***	.59***
ePortfolio: Knowledge of conventions	.43***	.67***
ePortfolio: Composing in electronic environments	.53***	.76***
ePortfolio: Holistic score	.53***	.77***

*** *p* < .001

significant levels (*t*[2959] = 20.536, *p* < .001). In enrolled and predictive patterns, students writing course grades in the first course (*n* = 2,172, *M* = 3.0, *SD* = 1.09, Range = 0, 4) and the second (*n* = 2,147, *M* = 3.11, *SD* = .96, Range = 0, 4) were high. In the case of ePortfolio holistic scores, the mean score of NJIT students (*n* = 210, *M* = 7.60, *SD* = 2.17, Range = 2, 12) is above the score of 7—the warning score that students may not be performing at agreed-upon levels of proficiency.

Reliability Evidence

As noted above, Haertel (2006) defined reliability in terms of replication. In the case of ePortfolio scores, questions of inter-reader reliability remain an important prerequisite to score interpretation and use. Important to interpretation of information presented in Tables 3 and 4 are distinctions by Stemler (2004) regarding consensus and consistency estimates.

In the case of UI, consensus estimates of inter-reader reliability are appropriate to the aim of certification of student ability. Based on the assumption that skilled readers should be able to come to exact agreement about how to apply various levels of a scoring rubric to an ePortfolio at hand, consensus estimate of inter-reader reliability are computed through the use of percent-agreement, as demonstrated in Table 3. On the left side of Table 3, each score level is identified from the highest (6) to lowest (1). Because certification is the assessment aim, categories are developed to

Table 5
University of Idaho Correlation of Criterion Measures: All Students

Measures	1	2	3	4	5	6
1. HS GPA (*N* = 1161)	—	.23**	.31**	.23**	.22**	.33**
2. SAT writing (*N* = 919)		—	.64**	.16**	.05	.12**
3. ACT composite (*N* = 594)			—	.07	-.02	.12**
4. ePortfolio score (*N* = 1208)				—	.82**	.15**
5. Eng.101 course grade *(N* = 1208)					—	.05
6. Eng. 102 course grade (*N* = 971)						—

*p < .05 **; p < .01

Table 6
NJIT Correlation of Criterion Measures: All Students

Measures	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
1. HS rank (*N* = 1420)	—	.30**	.12	.14	.27**	.23**	.23*	.20*	.33**	.32**
2. SAT writing (*N* = 2636)		—	.16*	.13	.13	.29**	.03	.16*	.24**	.27**
3. ePortfolio: Rhetorical knowledge (*N* = 210)			—	.84**	.60**	.71**	.59**	.84**	.10	.18*
4. ePortfolio: Critical thinking (*N* = 210)				—	.62**	.71**	.57**	.82**	.14*	.26**
5. ePortfolio: Writing processes (*N* = 210)					—	.61**	.50**	.70**	.14*	.24**
6. ePortfolio: Knowledge of conventions (*N* = 210)						—	.44**	.73**	.19**	.21**
7. ePortfolio: Composing in electronic environments (*N* = 210)							—	.69**	.05	.12
8. ePortfolio: Holistic score (*N* = 210)								—	.18*	.20**
9. Hum. 101 course grade (*N* = 2171)									—	.39**
10. Hum. 102 course grade (*N* = 2147)										—

**p* < .05; ***p* < .01

determine failure (due to absence of required materials or plagiarism), no pass, pass, and distinction. To assure deliberative review of student ePortfolios, program administrators have established two tiers of review using the expert-rater method of evaluation articulated by Smith (1993) and Haswell and Wyche (1996). While Tier 1 functions as an initial review, Tier 2 leverages second readings when adjudication is needed. While, for instance, the category of no pass is used to justify course failure, a second reading is required to substantiate that judgment. The same is true for the category of distinction. The rating methodology focuses the attention of the rater where most disagreement occurs: at either the low or high end of the evaluation scale. The ePortfolios that obviously are ready for the next first-year writing course are not read a second time. Efficacy of the model is shown by the Gaussian (normal) distribution on the right side of Table 3. With passing scores of 4 (*n* = 926, or 76.7% of the scores) at the apex of the bell curve, the two tails occur as expected with the higher (scores of 5 and 6) and lower (scores of 1 and 2) ends of the distribution.

In the case of the NJIT scores, inter-reader reliability—termed consistency estimated by Stemler (2004)—is determined for each variable. Scores are reported in Pearson product moment correlations to document their non-adjudicated and adjudicated forms. For example, a score that is matching (6 + 6) or adjacent (6 + 5) is not adjudicated; however, if a score is beyond adjacent (6 + 4), a third reader is called upon to reconcile the scores. That third score is used to determine the final score. So, if an ePortfolio receives a trait or holistic score of 6 by one reader and 4 by a second, and if the third reader gives it a score of 5, then the total score of 11 is awarded. If, however, the third reader awards a score of 3, then the total score is lowered and recorded as 7.

Reporting both non-adjudicated and adjudicated scores allows assessment stakeholders to determine the degree to which score consistency was reached. As Table 4 shows, statistically significant non-adjudicated Pearson correlations range from 0.42 to 0.53, a medium level of correlation. Under conditions of adjudication, scores rise as expected from a low of 0.59 to a high of 0.77, medium-to-high levels of correlation.

Validity Evidence

As noted above, Kane (2013) conceptualized validity through score interpretation and use. As part of the validity argument, correlation evidence is used to help stakeholders understand the relationship of ePortfolio scores to the pre-college measures, enrolled measures, and predictive measures identified in Tables 1 and 2. Relying on this empirical tradition, in Table 5 we provide relational evidence from UI. We identified a statistically significant low correlation between high school GPA and all other measures. Also evident is a moderate relationship between the SAT Writing and ACT Composite scores. There is a high correlation between the ePortfolio score and the writing course grade. However, correlation between the ePortfolio score and the next writing course is low—and there is no statistically significant relationship between writing courses.

Table 6 provides similar relational evidence from NJIT. High school rank demonstrates a statistically significant low correlation with all measures. SAT Writing scores correlate at statistically significant low levels with present and next writing course grade. Trait and ePortfolio holistic scores correlate at statistically significant medium-to-high levels. Correlation of the ePortfolio holistic score and present writing course grade is low, though statistically significant. While statistically significant, correlation between the holistic score and the next writing course is also low, and correlation between present and next writing course is medium.

Fairness Evidence

Were we to stop the analysis here, with only the most general trends, we would find our evidence related to reliability and validity useful but limited. Categories of evidence are deepened, however, when fairness is centralized in the analysis. To this end, we conducted a brief disaggregated analysis at NJIT to demonstrate the need for detailed sub-group information. More in-depth attention is given to demographic characteristics at both UI and NJIT

Disaggregated reliability consensus estimates. The importance of disaggregating reliability information according to student sub-groups is illustrated in the NJIT ePortfolio data comparing the overall population and female students (Figures 3 and 4). As Figure 3 illustrates, non-adjudicated scores for all students shown in Table 4 ranged from 0.37 (writing processes) to 0.54 (critical thinking). However, scores for female students, as Figure 3 shows, are much lower, ranging from a low of -0.04 on writing processes to a high of 0.44 on the holistic score. While not shown in Figure 3, only the holistic score achieved statistical significance. As Figure 4 illustrates, scores for the overall population improved upon adjudication, ranging from 0.59 (writing processes) to 0.77 (holistic score). As Table 4 shows, reader scores correlated at statistically significant levels ($p < .001$). Yet, Figure 4 also demonstrates that the adjudicated scores were low for female students, ranging from 0.02 (*nss*) to 0.63 ($p < .001$). Based on the disaggregated information shown in Figures 3 and 4, a radically different picture of consistency appears for female students.

Disaggregated student profiles. Returning to Table 1 at UI, attention is given to gender, race/ethnicity, first-generation college status, Pell grant status, and EFC quartiles in the UI study. An analysis of students' EFC levels from the Idaho State Longitudinal Data System SLDS was recoded to match the family income levels listed in the State Profile Report for college bound seniors in Idaho (College Board, 2015a, p. 4). The College Board listed ten income categories, and then the UI data was recoded into quartiles for EFC analysis. This process allowed for a reasonable and representative portrait of students' family income levels at the UI.

Statistically significant differences are noted for all pre-college measures between male and female students. Depending on measure, sub-group differences are noted for all except first generation, Pell Grant, and second quartile of EFC students. In terms of enrolled college measures, statistically significant differences are present only between the ePortfolio holistic scores of men and women and between first generation students and other sub-groups. In terms of writing course grade, statistically significant differences appear only between male and female students. No statistically significant differences appear on next semester course grades.

Disaggregated validity correlations. To continue our disaggregated analysis with an expansion of Table 5, we provide details in Table 7 on the correlation information for UI students by first generation and Pell grant status. Table 7 demonstrates that the patterns for both categories of students are similar—high school GPAs correlate at a low but statistically significant

Figure 3

NJIT Non-Adjudicated Pearson Correlation Coefficients, All Students and Female Students

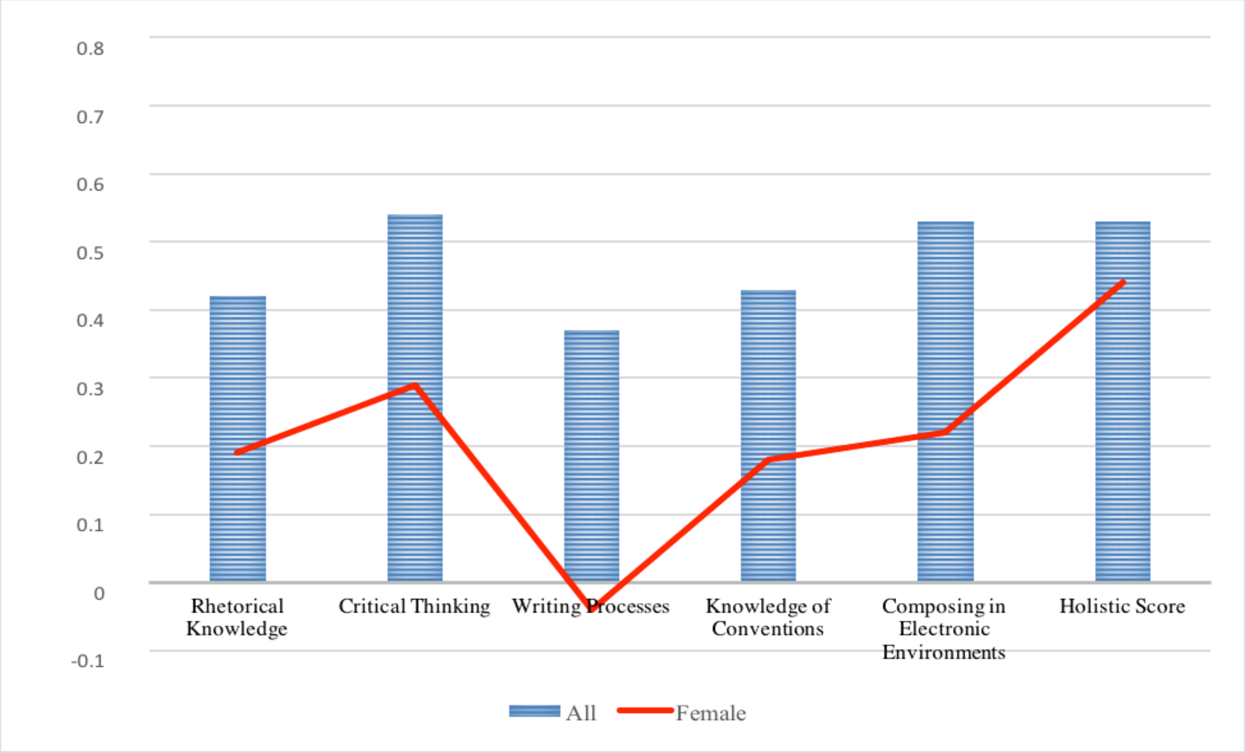
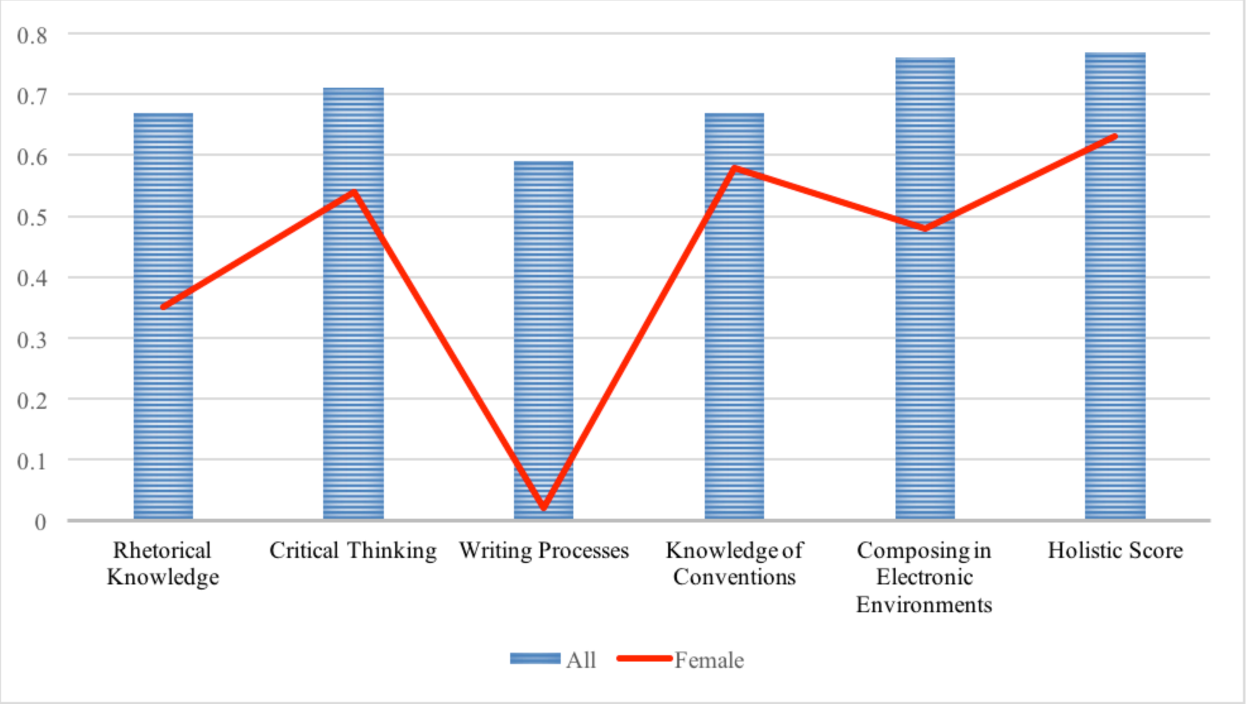


Figure 4

NJIT Adjudicated Pearson Correlation Coefficients, All Students and Female Students



level across all measures; holistic ePortfolio scores and writing course grades correlate at a high statistically significant level; and next writing course grade has no statistically significant relationship to ePortfolio holistic score.

Tables 8 and 9 contain correlations among pre-college, concurrent, and predictive measures disaggregated by EFC status at the University of Idaho from the Idaho SLDS database. Similar patterns to the other measures are observed across all EFC groups. That is, although the descriptive statistics in Table 1 showed statistically significant differences between the first and second EFC quartile and the third and fourth quartiles, the disaggregated pre-college measures and enrolled college measures have similar correlations and strength. As in Table 7, Tables 8 and 9 show statistically significant low correlations between high school GPA and all other measures; a medium to high correlation between the SAT Writing and ACT Composite scores; and a strong relationship between the ePortfolio score and the writing course grade. The lack of a statistically significant relationship between the first and second writing courses remains across EFC groups, and the ePortfolio holistic score maintains its strongest statistically significant relationship with first writing course grades. Although low, statistically significant relationships are maintained across all EFC groups with ePortfolio scores and high school GPA. ePortfolio scores, however, demonstrate no statistically significant relationship to grades achieved in the second writing course.

Disaggregated predictive evidence. Regarding predictive evidence disaggregated by demographic characteristics, Table 9 provides information about the power of criterion measures to forecast writing measures at UI. With the exception of Hispanic students, the pre-college measures achieve statistical significance but account for no more than 14% of the variance (in Model 1A for students of two or more races) in their relationship to ePortfolio scores. Again, with the exception of Hispanic students, in terms of predicting writing course grade, pre-college measures achieve statistical significance but account for, at best, 17% of the variance (in Model 1C for students of two or more races). In their ability to predict writing course grade, ePortfolio scores achieve statistical significance for all student sub-groups under Model 2A, accounting for 65% of the variance for female and white students to 81% for students of two or more races. In terms of predictive ability for the second writing course, Model 3A accounts, at best, for 29% of the variance for male students; the model fails to achieve statistical significance for female, Hispanic students, and students of two or more races.

Table 11 provides information regarding the power of criterion measures to forecast writing measures at NJIT. Pre-enrollment measures identified in Model 1A fail to achieve statistical significance in terms of predicting the ePortfolio holistic score for the overall group and for all sub-groups. In predicting the writing course grade, statistical significance is achieved for the overall group and for all sub-groups, with the highest prediction for Asian students accounting for 16% of the variance. Enrollment measures in Model 2A achieve statistical significance for the overall group and for all sub-groups, accounting for 84% of the variance for male students. Model 2B fails to achieve statistical significance for female, Asian, and Hispanic students and, at best, accounts for 19% of the variance for white students. Model 3A, designed to predict the second writing course grade, achieved statistical significance for the overall group and for all sub-groups, with 34% of the variance accounted for Asian students.

Table 7

University of Idaho ePortfolio Score Correlations: First Generation and Pell Grant Status

Measures	1.	2.	3.	4.	5.	6.
1. HS GPA (First Gen *N* = 434; Pell *N* = 509)	—	.28***	.31**	.19**	.16**	.33**
2. SAT writing (First Gen *N* = 354; Pell *N* = 414)	.22**	—	.67**	.18**	.08	.08
3. ACT composite (First Gen *N* = 231; Pell *N* = 275)	.26**	.68**	—	.06	-.04	.10
4. ePortfolio holistic score (First Gen *N* = 441; Pell *N* = 516)	.24**	.13**	.11	—	.82**	.09
5. Eng. 101 course grade (First Gen *N* = 441; Pell *N* = 516)	.24**	.02	.02	.80**	—	-.00
6. Eng. 102 course grade (First Gen *N* = 360; Pell *N* = 416)	.33**	.10	.08	.15**	.05	—

Note. Correlations for first generation status are in the upper diagonal of the matrix and correlations for Pell Grant status are in the lower diagonal of the matrix.

**p* < .05; ***p* < .01

Table 8

University of Idaho ePortfolio Score Correlations: EFC Quartiles 1 and 2

Measures	1.	2.	3.	4.	5.	6.
1. HS GPA (EFC Q1 *N* = 313; EFC Q2 *N* = 253)	—	.13*	.17*	.25**	.26**	.29**
2. SAT writing (EFC Q1 *N* = 252; EFC Q2 *N* = 203)	.32**	—	.63**	.10	-.05	.08
3. ACT composite (EFC Q1 *N* = 169; EFC Q2 *N* = 137)	.36**	.71**	—	.01	-.10	.10
4. ePortfolio holistic score (EFC Q1 *N* = 313; EFC Q2 *N* = 253)	.18**	.20**	.20*	—	.80**	.09
5. Eng. 101 course grade (EFC Q1 *N* = 313; EFC Q2 *N* = 253)	.17**	.13	.17*	.79**	—	.03
6. Eng. 102 course grade (EFC Q1 *N* = 250; EFC Q2 *N* = 199)	.47**	.14	.08	.19**	.13	—

Note. Correlations for Q1 are in the upper diagonal of the matrix and correlations for Q2 are in the lower diagonal of the matrix. EFC Q1 = $0-$20,000 family income; EFC Q2 = $20,000-$60,000.
**p* < .05; ***p* < .01

Table 9

University of Idaho ePortfolio Score Correlations: EFC Quartiles 3 and 4

Measures	1.	2.	3.	4.	5.	6.
1. HS GPA (EFC Q3 *N* = 223; EFC Q4 *N* = 210)	—	.23**	.40**	.34**	.29**	.42**
2. SAT writing (EFC Q3 *N* = 185; EFC Q4 *N* = 179)	.20**	—	.60**	.18*	.07	.20*
3. ACT composite (EFC Q3 *N* = 116; EFC Q4 *N* = 110)	.28**	.62**	—	-.03	-.09	.32**
4. ePortfolio holistic score (EFC Q3 *N* = 226; EFC Q4 *N* = 211)	.26**	.14	.15	—	.84**	.15*
5. Eng. 101 course grade (EFC Q3 *N* = 226; EFC Q4 *N* = 211)	.23**	.10	.04	.82**	—	.00
6. Eng. 102 course grade (EFC Q3 *N* = 176; EFC Q4 *N* = 186)	.31**	.09	.10	.10	-.05	—

Note. Correlations for Q3 are in the upper diagonal of the matrix and correlations for Q4 are in the lower diagonal of the matrix. EFC Q3 = $60,000-$100,000; EFC Q4 = $100,000+.
**p* < .05; ***p* < .01

Discussion

We frame our comments in terms of our three research questions to discuss the application of empirical methods to ePortfolio-based assessment at two distinctly different universities with selective undergraduate student profiles.

Reliability Evidence: Consensus and Consistency Estimates

In addressing the relationship between reliability and validity, Mislevy (2004) asserted that researchers must not sell techniques short based on standard practice. To do so is to "miss the compiled wisdom underlying those techniques" (Mislevy, 2004, p. 244). In the examples presented in Tables 3 and 4, we demonstrate that there are multiple ways to conceptualize, execute, and present information on inter-reader reliability. Indeed, as a way to conceptualize inter-reader reliability, the model offered by Stemler (2004) provides a straightforward method to attack complex evidentiary problems related to precision.

In terms of inferences based on this information, we conclude that the holistic method used at UI is well

Table 10

University of Idaho Regression Models: Pre-College Enrollment Measures, Enrolled Measures, Predictive Measeures by Gender and Race/Ethnicity

	Pre-College Enrollment Measures								Enrolled College Measures			
									Concurrent		Predictive	
	Model 1A HSGPA + SAT writing→ ePortfolio holistic score		Model 1B HSGPA + ACT composite→ ePortfolio holistic score		Model 1C HSGPA + SAT writing→ Eng. 101 course grade		Model 1D HSGPA + ACT composite→ Eng. 101 course grade		Model 2A ePortfolio score→ Eng. 101 course grade		Model 3A ePortfolio score + Writing course grade → Eng. 102 course grade	
	R^2	F	R^2	F	R^2	F	R^2	F	R^2	F	R	F
All	.087	$F(2, 911) =$ 44.423***	.055	$F(2, 583) =$ 6.879***	.066	$F(2, 911) =$ 32.230***	.053	$F(2, 583) =$ 16.318***	.667	$F(1, 2015) =$ 2416.52***	.021	$F(2, 968) =$ 10.596***
Male	.083	$F(2, 461) =$ 20.918***	.063	$F(2, 312) =$ 10.450***	.074	$F(2, 461) =$ 18.411***	.057	$F(2, 312) =$ 9.370***	.673	$F(1, 648) =$ 1337.70***	.029	$F(2, 498) =$ 7.423**
Female	.075	$F(2, 447) =$ 18.099***	.030	$F(2, 268) =$ 4.213*	.040	$F(2, 447) =$ 9.284***	.033	$F(2, 268) =$ 4.550*	.652	$F(1,556) =$ 1040.10***	.011	$F(2, 467) =$ 2.544 *nss*
White	.072	$F(2, 671) =$ 26.077***	.042	$F(2, 439) =$ 9.720***	.058	$F(2, 671) =$ 20.707***	.054	$F(2, 439) =$ 12.510***	.652	$F(1, 858) =$ 1606.39***	.015	$F(2, 690) =$ 5.109**
Hispanic	.125	$F(2, 125) =$ 8.947***	.057	$F(2, 82) =$ 2.470 *nss*	.083	$F(2, 125) =$ 5.668**	.023	$F(2, 82) =$ .966 *nss*	.665	$F(1, 151) =$ 299.892***	.022	$F(2, 129) =$ 1.421 *ns*
Two or More Races	.140	$F(2, 40) =$ 3.265*		*qns*	.165	$F(2, 40) =$ 3.942*		*qns*	.809	$F(1, 48) =$ 203.59***	.033	$F(2, 40) =$ .676 *nss*

Note. p values not statistically significant at the 0.05 level are designated as *nss*. Sample sizes under 30 are designated as *qns*.
$^*p < .05$; $^{**}p < .01$; $^{***}p < .001$.

Table 11

NJIT Regression Models: Pre-College Enrollment Measures, Enrolled Measures, Predictive Measures by Gender and Race/Ethnicity

	Pre-College Enrollment Measures				Enrolled College Measures					
					Concurrent				Predictive	
	Model 1A HS rank + SAT writing→ holistic score		Model 1B HS rank + SAT writing→ Hum. 101 course grade		Model 2A ePortfolio Traits→ ePortfolio holistic score		Model 2B ePortfolio Traits + ePortfolio holistic score→ Hum. 101 course grade		Model 3A ePortfolio traits + ePortfolio holistic score + Writing course grade→ Hum. 102 course grade	
	R^2	F	R^2	F	R^2	F	R^2	F	R^2	F
All	.046	$F(2, 109) =$ 2.6 *nss*	.139	$F(2, 1077) =$ 88.16***	.83	$F(5, 204) =$ 199.16***	.06	$F(6, 200) =$ 2.13*	.232	$F(7, 177) =$ 7.65***
Male	.026	$F(2, 96) =$ 1.25 *nss*	.119	$F(2, 881) =$ 59.53***	.838	$F(5, 173) =$ 179.59***	.079	$F(6, 170) =$ 2.44*	.252	$F(7, 151) =$ 7.26***
Female		*Qns*	.25	$F(2, 193) =$ 30.84***	.766	$F(5, 25) =$ 16.34***	.088	$F(6, 23) =$ .369 *nss*		*Qns*
White	.038	$F(2, 48) =$ .955 *nss*	.154	$F(2, 404) =$ 36.83***	.824	$F(5, 83) =$ 77.73***	.192	$F(6, 81) =$ 3.20**	.317	$F(7, 71) =$ 4.70***
Asian		*qns*	.156	$F(2, 216) =$ 21.16***	.819	$F(5, 53) =$ 47.89***	.04	$F(6, 51) =$ .325 *nss*	.344	$F(7, 44) =$ 3.29**
Hispanic		*qns*	.106	$F(2, 244) =$ 15.59***	.80	$F(5, 30) =$ 24.20***	.116	$F(6, 28) =$ .611 *nss*		*qns*
Black		*qns*	.16	$F(2, 107) =$ 10.17***		*qns*		*qns*		*qns*

Note. p values not statistically significant at the 0.05 level are designated as *nss*. Sample sizes under 30 are designated as *qns*.
$^*p < .05$; $^{**}p < .01$; $^{***}p < .001$

suited to certification assessment aims in which each ePortfolio has to be read. Reference to consensus methods and Gaussian distribution provides additional evidence that the scores are normally distributed and that, in cases of discrepancy, measures are in place to ensure adjudication. We may also conclude that the consistency reliability measures at NJIT are well suited to the aim of program assessment in which multiple traits provide information that can, in turn, be used to structure opportunities to learn for students by curricular refinement.

However, this is not to say that positive claims are free from qualification. At UI, the single ePortfolio holistic score is just that—a single score upon which a judgment is to be made. While the multiple-trait method used at NJIT may appear preferable, that method would take additional time and resources; furthermore, it is not clear what role the traits would serve in a certification assessment.

In terms of disaggregation by sub-group illustrated in Figures 3 and 4, none of the non-adjudicated scores for females reached levels of statistical significance, and even under adjudication the writing processes scores failed the test of statistical significance. In stark contrast to the 0.77 holistic score inter-reader reliability reported for the overall population, consistency estimated achieved only a moderate .48 (p < .01) for female students. In terms of score interpretation and use, it would be difficult to justify the use of scores for any purpose regarding inferences about the writing ability of female students at NJIT based on ePortfolio scores. In the case of inter-reader reliability evidence related to ePortfolio scoring, investigating complex evidentiary problems related to precision may result in important reservations about score interpretation and use.

Validity Evidence: Correlation Analysis

At UI and NJIT, ePortfolio scores had a demonstrable place in the writing ecology at the institution, with the scores appropriately interpreted in relationship to curricular aims. Any institution, in fact, would benefit from the level of information associated with the two case studies.

Positive claims involving validity evidence are accompanied by qualifications related to assessment purpose. At UI, ePortfolio scores are used to certify students and, as such, the measure is not entirely independent of course grade. This interdependence helps to explain the high, statistically significant correlation between ePortfolio score and Eng. 101 course grade (r = .82, p < .01) shown in Table 5 and persisting with little variation across sub-groups in Tables 7 through 9. Used for a different purpose at NJIT, ePortfolio scores used for program assessment are independent of course grade. Table 6 illustrates the statistically significant, low correlation between ePortfolio score and Hum. 101 course grade (r = .18, p < .05). At NJIT, the ePortfolio scores demonstrate disjuncture between the average course grade of B shown in Table 2 (M = 3.0, SD = 1.09) and ePortfolio trait and holistic scores well below the ranges associated with above average work.

In both institutions, the relationship of scores to subsequent course writing grades was weaker than anticipated. As Table 1 reveals, at UI the average grade in the second writing course is higher, at statistically significant levels, than the first writing course (t[1336] = 10.32, p < .001); furthermore, 20% of the students are lost from the first course to the second. At NJIT, the absence of relationship is due to relatively low trait and holistic scores compared to course grades of B in the second semester writing course, as well as the first. Such evidence reveals the need for connections between assessment findings and curricular refinement. In the case of both institutions, there is evidence that across-course ePortfolio development is worth consideration in order to examine relationships between assessment scores and other forms of convergent evidence. As Elliot et al. (2016) have noted, attention to ePortfolio scores in relationship to criterion measures of the writing construct allows detailed information to be obtained on study-site ecologies—including ways that new digital forms of assessment mediate the writing construct and may inadvertently result in construct underrepresentation.

In terms of reservations regarding validity evidence, it is also worth recalling that even the most robust assessments cannot hope to capture the writing construct in its entirety. Writing instruction and writing assessment are best conceptualized by attending to cognitive, interpersonal, intrapersonal, and physiologic domains. Only an expert instructor observing students over long periods of time is qualified to make inferences about an individual student ability in these domains. No ePortfolio-based assessment evaluated in a scoring session, whether by holistic or trait methods, can hope to capture full representation of the writing construct. To begin with this premise is to appreciate the strengths and limits of ePortfolio assessment.

Fairness: Statistically Significant Difference and Regression Analysis

Portraits of students presented in Tables 1, 2, 7, 8, 9, 10, and 11—as well as in Figures 3 and 4—afford a deservedly complex view of how various students perform by demographic category, compared to the aggregated portrait that includes all students. While some sub-group categories are familiar, others presented in the UI case study are new. Recently, for example, emerging research documented that there was little agreement on how first-generation students were defined, but "regardless of how they're defined, first-generation students enroll and graduate at lower rates than do other students" (Smith, 2015, para. 4) and are thus a group of interest.

In terms of evidence related to fairness, ePortfolio scores predict the first writing course grade across all gender and race/ethnicity groups, with prediction at its highest for students of two or more races. While not shown in the present study, this pattern persists across first-generation, Pell grant, and EFC students (at rates no lower than 63%). Minimum group differences in ePortfolio scores among race/ethnicity groups in Table 1 bring our emphasis on principles of fairness full circle. And, while statistically significant group means are identified between men and female and between first generation and non-first generation students, no statistically significant differences are present among Pell grant and EFC students. Absence of group difference in ePortfolio scores is similar at NJIT. While there are indeed differences between males and females, no differences were observed among White, Asian, and Hispanic students (*F*[3, 192] = 1.82, *p* = .14). On either campus, this is a claim that cannot be substantiated for standardized measures such as the SAT Writing section where statistically significant group differences are everywhere apparent.

The absence of group differences in ePortfolio scores leads us to hypothesize one of the most important findings of the study: Robust construct representation leads to fairness in writing assessment; conversely, constrained construct representation leads to group differences. ePortfolios have been touted for their flexibility across learning environments, and our study suggests that they are also flexible in their accommodation of learning demonstrated by diverse learners. While group differences in standardized measures such as the SAT Writing may lead to disparate impact—unintended racial differences in outcomes resulting from facially neutral policies or practices that on the surface seem neutral but nevertheless have the same consequence as overt discrimination (Poe, Elliot, Cogan, & Nurudeen, 2014)—our results suggest ePortfolios may be a way to minimize this negative impact.

In terms of reservations regarding evidence related to fairness, our analysis also highlights that students may concurrently occupy demographic spaces that place them in positions of both advantage and disadvantage. What actions, for example, do we take in realizing that female students outperform male students in writing ability? Are these tasks that serve some student groups better than others in advancing opportunity to learn? While a better understanding of student differences must qualify any claim of fairness, the results presented in the two case studies reveal new problems for us to solve. Such analysis encourages us to think about such complexities in our assessment reporting and to move beyond categorization of our students in isolated, demographic silos. Once we can begin to understand how student characteristics interact with domains of writing, we can then begin to chart an equitable and just way forward.

Conclusion

We want to close by proposing guiding questions for quantitative reporting of information related to ePortfolio score interpretation and use. To that end, we offer the questions in Table 12 through the three foundational measurement concepts of fairness, reliability, and validity in order to guide future practice related to ePortfolio score use and interpretation. Under the integrative principle of fairness and its association with opportunity to learn, dividing the table into questions of resource allocation and stakeholder lends specificity to the question: What do the empirical study results *mean* in terms of score interpretation and use? In other words, instead of focusing on the interpretation and use of ePortfolio scores to maintain course quality (at UI) and strengthen program assessment (at NJIT), we reconceptualize these aims as instrumental and therefore secondary to the advancement of opportunity to learn. The primary aim, advancement of opportunity to learn, subsumes all other assessment aims and compels us to reflect on the learning environment, demand articulated connections between the assessment and the instructional environment, and provide resources for the least advantaged students.

To achieve the dual aim of integrative and principled action identified in Table 12, administrators are invited to use ePortfolio scores for traditional aims—such as the maintenance of course quality and enhancement of program assessment—but these aims are restructured to include improvement of the learning environments for those students who appear to be least advantaged. Returning to Tables 1 and 2, administrators would allocate resources to further investigation of the ePortfolios themselves to determine why females score higher than men and why the scores of first generation students differ from those whose parents attended college.

Returning to Figure 1 and 2, administrators would also allocate resources to discover why the ePortfolios of NJIT female students—whose scores are higher than male students—resulted in rating complexities. As Moss (2004) had recommended, here is an excellent opportunity to use qualitative analysis in order to understand contradictory information. In practical

Table 12
Guiding Questions: Evidential Quantitative Reporting for ePortfolios

Standard	Infrastructure resources	Students	Instructors	Administrators (departmental & institutional)	Workforce
Fairness: "All steps in the testing process, including test design, validation, development, administration, and scoring procedures, should be designed in such a manner as to minimize construct-irrelevant variance and to promote valid score interpretations for the indented used of all examinees in the intended population" (Standard 3.0, p. 63).	How does the institution identify and support opportunity to learn for all students through ePortfolio assessment?	What type of evidence documents sensitivity to diverse ways of making meaning in ePortfolio assessment?	How do teachers ensure that diverse learners have opportunity to learn the construct assessed in ePortfolios?	How can administrators use ePortfolio findings to improve learning environments for the least advantaged students?	How can ePortfolio scores support opportunity to learn for all students beyond graduation and into workplace settings?
Reliability: "Appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use" (Standard 2.0, p. 42).	How do institutional technological resources and articulated program outcomes support a stable evaluation of student work?	What does score disaggregation reveal about inter-reader agreement and inter-reader reliability patterns of student groups?	How do the estimates of reliability influence score interpretation and use?	What is the evidence that the administrative and working conditions of the ePortfolio assessment have remained stable?	How are estimates of reliability determined across settings in terms of writing task demands?
Validity: "Clear articulation of each intended test score should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided" (Standard 1.0. p. 23).	How has the institution leveraged ePortfolio assessment to ensure robust construct representation for all students?	How are scores used to draw inferences about students' abilities assessed through ePortfolios, and how are these inferences interpreted?	How is teacher knowledge used in making decisions about individual student performance in ePortfolio assessment?	What kind of validity evidence is necessary to support principled interpretation and responsible score use of ePortfolios for multiple uses?	How are construct validity demands rendered congruent across academic and workplace settings?

Note. Fairness: Integrative and principled action

terms, examination of the scores of female students is an opportunity for researchers to examine the design of the ePortfolios themselves to see how members of this student group featured their skills in meeting course objectives. As well, examination of the ePortfolio scores of female students would examine the possibility of incorrect consistency estimates resulting in low correlations due to small sample size and range restrictions. For example, students who featured digital artifacts (e.g., blogs, social networking, video sharing, and wikis) may have elicited a wide range of discrepant scores if instructors were not accustomed to evaluating such artifacts; however, in a holistic score, these same ePortfolios may have received high scores. Only in-depth qualitative analysis would identify such patterns that could, in turn, be used to help all students design their ePortfolios with greater audience awareness.

Along with infrastructure resources determined by administrators, Table 10 calls attention to the importance of the validity inferences made about students. While there is a longstanding tradition in writing studies of distinguishing between low stakes and high stakes writing (Elbow, 1997), attention to fairness helps us to realize that all interpretations and inferences we make about our students are of great consequence. The uses we make of performance scores are all high stakes because they embody impressions of student ability. While, for example, Model 1A and Model 1C illustrate statistical significance of high school GPA and SAT Writing in predicting, respectively, ePortfolio scores and course grades across all groups, these models cover so little of the variance (no more than 14% for students of two or more races) that questions arise regarding the use of these models for any interpretation whatsoever, including admissions and placement purposes. In similar fashion, comparison of Model 2A and Model 3A suggests that ePortfolio scores are most useful when they are aligned to specific courses and of less value across courses. In terms of impact on students and the inferences we make about them, emphasis on opportunity to learn compels us to realize that qualifications must be drawn across all assessments, regardless of the degree of construct representation. While ePortfolios are often understood as "an antidote to the inadequacies of testing" (Cambridge, Cambridge, & Yancey, 2009, p. 195), their perceived face validity does not negate the need for justification of their use and qualification of their limits in the inferences we draw about student ability.

While his focus is on tests of language, Cumming (2013) emphasized that integrated writing tasks focus on uses of written language to construct knowledge, often in multimodal ways, which involve genres that are ill-defined and so difficult to score. Accustomed to a print environment in high school, many students, among them those at NJIT, struggled when faced with new genres—as did their instructors, who had used the source-based essay as the exclusive reporting structure in first-year writing. It is therefore important to remember that the ability to achieve proficiency in these new genres is compounded if there are any student weaknesses in writing ability in the first place. In their study of the digital skills of 91 low-income students enrolled in writing remediation, Relles and Tierney (2014) found that students who are underprepared according to traditional writing criteria face additional barriers to academic success because of low digital skills. "Today's remedial writers," they concluded, "may be challenged by a kind of literacy double jeopardy that is unique to the 21st century" (Relles & Tiernet, 2014, p. 497). In the classroom, instructors may be especially challenged to ensure that students have both the traditional and digital abilities to prepare the integrated writing tasks that are often part of the new genre of ePortfolios themselves.

In closing, we want to call attention again to the contention by Yancey at al. (2013) that ePortfolio assessment requires a new vocabulary and a new set of practices. We agree, and our work here is intended to contribute to the role that empirical assessment should play in such new theoretical models. While the techniques we have illustrated are traditional, emphasis on fairness as vehicle for integrative, principled action intended to advance opportunity to learn is unique. While conceptual advantages have been presented here in terms of ePortfolio score interpretation and use, additional work will be needed if empirical and theoretical domains are to function in complementary fashion in order to structure opportunity for students.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: APA.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

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

Bazerman, C. (2015). What do sociocultural studies of writing tell us about learning to write? In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed., pp. 24-40). New York, NY: Guilford.

Behizadeh, N., & Englehard, G. (2015). Valid writing assessment from the perspectives of the writing and measurement communities. *Pensamiento Educativo: Revista de Investigación Educacional Latinoamericana, 52*, 34-54.

Belanoff, P., & Elbow, P. (1986). Using portfolios to increase collaboration and community in a writing program. *Writing Program Administration, 9*, 27-40.

Brennan, R. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.

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

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

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

Charney, D. (1996). Empiricism is not a four-letter word. *College Composition and Communication, 47*, 567-593. doi:10.2307/358602

College Board. (2015a). *State profile report: Idaho*. Retrieved from https://secure-media.collegeboard.org/digitalServices/pdf/sat/ID_15_03_03_01.pdf

College Board. (2015b). *State profile report: New Jersey*. Retrieved from https://secure-media.collegeboard.org/digitalServices/pdf/sat/NJ_15_03_03_01.pdf

Coppola, N., & Elliot, N. (2010). Assessment of graduate programs in technical communication. In M. Hundleby & J. Allen (Eds.), *Assessment in technical and professional communication* (pp. 127-160). New York, NY: Baywood.

Council of Writing Program Administrators, National Council of Teachers of English, & National Writing Project. (2011). *Framework for success in postsecondary writing*. Retrieved from http://wpacouncil.org/files/framework-for-success-postsecondary-writing.pdf

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly, 10(1),* 1-8. doi:10.1080/15434303.2011.622016

Elbow, P. (1986). State University of New York, Stony Brook: Portfolio-based evaluation program. In. P. Connolly & T. Vilardi (Eds.), *New methods in college writing programs: Theory into practice* (pp. 95-105). New York, NY: Modern Language Association.

Elbow, P. (1997). High stakes and low stakes in assigning and responding to writing. *New Directions for Teaching and Learning, 69*, 5-13. doi:10.1002/tl.6901

Elliot, N. (2005). *On a scale: A social history of writing assessment in America*. New York, NY: Lang.

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

Gee, J. P. (2008). A sociocultural perspective on 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. 76-108). Cambridge, UK: Cambridge University Press.

Grossman, L. (2006, December 12). Time person of the year: You. *Time Magazine, 68*(126), pp. 38-41.

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

Hamp-Lyons, L. (2016). Editorial: Farewell to holistic scoring? *Assessing Writing, 27*, A1-A2. doi:10.1016/j.asw.2016.06.006

Haswell, R. H., & Wyche, S. (1996). A two-tiered rating procedure for placement essays. In T. W. Banta, J. P. Lund, K. E. Black, & F. W. Oblander (Eds.), *Assessment in practice: Putting principles to work on college campuses* (pp. 204-207). San Francisco, CA: Jossey-Bass.

Johnson, C. S., & Elliot, N. (2010). Undergraduate technical writing assessment: A model. *Programmatic Perspectives, 2*, 110-151. Retrieved from http://www.cptsc.org/pp/vol2-2/johnson_elliot2-2.pdf

Kane, M. T. (2013). Validating the interpretation and uses of test scores. *Journal of Educational Measurement, 50*, 1-73. doi:10.1111/jedm.12000

Kelly-Riley, D. (2012). Setting sail with Ed White: The possibilities of assessment and instruction within college writing assessment. In N. Elliot & L. Perelman (Eds.), *Writing assessment in the twenty-first century: Essays in honor of Edward M. White* (pp. 157-168). New York, NY: Hampton Press.

Kelly-Riley, D. and Whithaus, C., eds. (2016) Special issue on a theory of ethics for writing assessment, *Journal of Writing Assessment, 9*(1) http://journalofwritingassessment.org.

Kerlinger, F. N., & Lee, H. B. (1999). Foundations of behavioral research. Mason, OH: Cengage Learning.

Ketcheson, K. A. (2009). Sustaining change through student, departmental, and institutional portfolios. In D. Cambridge, B. Cambridge, & K. Yancey (Eds.), *Electronic portfolios 2.0: Emergent research on implementation and impact* (pp. 137-143). Serling, VA: Stylus.

Klobucar, A., Elliot, N., Deess, P., Rudniy, O., & Joshi, K. (2013). Automated scoring in context: Rapid assessment for placed students. *Assessing Writing, 18*(1), 62-84. doi:10.1016/j.asw.2012.10.001

Larkin, M. L., & Robertson, R. L. (2013). Complex moving parts: Assessment system and electronic portfolios. *International Journal of ePortfolio, 3*, 27-37. Retrieved from http://www.theijep.com/pdf/ijep89.pdf

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012-1027. doi:10.1037/0003-066X.35.11.1012

Middle States Commission on Higher Education. (2015, August). *Newsletter: Special edition: Middle States Commission on Higher Education.* Retrieved from http://www.msche.org/newsletters/August-2015-Special-Edition150818182808.pdf

Mislevy, R. J. (2004). Can there be reliability without "reliability?" *Journal of Educational and Behavioral Statistics, 29*, 241-244. doi:10.3102/10769986029002241

Mislevy, R. J. (2016). How developments in psychology and technology challenge validity argumentation. *Journal of Educational Measurement, 53*, 1-19. doi:10.1111/jedm.12117

Moss, P. A. (2004). The meaning and consequences of "reliability." *Journal of Educational and Behavioral Statistics, 29*, 245-249. doi:10.3102/10769986029002245

Onink, T. (2014. November 28). 2015 guide to FAFSA, CSS profile, college aid and expected family contribution (EFC). *Forbes*. Retrieved from http://www.forbes.com/sites/troyonink/2014/11/28/2015-guide-to-fafsa-css-profile-college-financial-aid-and-expected-family-contribution-efc/

Organization for Economic Co-Operation and Development. (2014). *Measuring the digital economy: A new perspective.* Paris, France: OECD. Retrieved from http://www.oecd-ilibrary.org/science-and-technology/measuring-the-digital-economy_9789264221796-en

Phelps, L. W., & Ackerman, J. W. (2010). Making the case for disciplinarity in rhetoric, composition, and writing studies: The visibility project. *College Composition and Communication, 62*, 180-215.

Poe, M., Elliot, N., Cogan, J. A., & Nurudeen, T. G. (2014). The legal and the local: Using disparate impact analysis to understand the consequences of writing assessment. *College Composition and Communication, 65*, 588-611.

Pullin, D. C. (2008). Assessment, equity, and 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. 333- 351). Cambridge, UK: Cambridge University Press.

Relles, S. R., & Tierney, W. G. (2013). Understanding the writing habits of tomorrow's students: Technology and college readiness. *Journal of Higher Education, 84*, 477-505. doi:10.1353/jhe.2013.0025

Rhodes, T. (2011, January). Making learning visible and meaningful through electronic portfolios. *Change: The Magazine of Higher Learning*. Retrieved from http://www.changemag.org/archives/back%20issues/2011/january-february%202011/making-learning-visible-full.html

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

Smith, A. A. (2015, November 10). Who are first generation students and how do they fare? *Inside Higher Ed*. Retrieved from https://www.insidehighered.com/news/2015/11/10/who-are-first-generation-students-and-how-do-they-fare

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

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

Suskie, L. (2009). *Assessing student learning: A common sense guide*. San Francisco, CA: Jossey-Bass.

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

Williams, J. M. (2010). Evaluating what students know: Using the RoseE Portfolio system for institutional and program outcomes assessment tutorial. *IEEE Transactions on Professional Communication, 53*, 46-57. doi:10.1109/TPC.2009.2038737

Yancey, K. B. (2009). Electronic portfolios a decade into the twenty-first century: What we know, what we need to know. *Peer Review, 11*, 28-92.

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 and evaluation*. Logan, UT: Computers and Composition Digital Press. Retrieved from http://ccdigitalpress.org/dwae/08_yancey.html

---

DIANE KELLY RILEY is Associate Professor of English and Director of Writing at the University of Idaho.

NORBERT ELLIOT is Professor Emeritus of English at New Jersey Institute of Technology. In 2016, he was appointed Research Professor at the University of South Florida.

ALEX RUDNIY is Assistant Professor of Computer Science at Fairleigh Dickinson University. From 2012 to 2014, he was Data Manager at the Office of Institutional Research and Planning at New Jersey Institute of Technology.