Jl. of Interactive Learning Research (2013) **24**(1), 417-444

Web-Based Portfolio Assessment: Validation of an Open Source Platform

REGINA COLLINS, NORBERT ELLIOT,
ANDREW KLOBUCAR, AND FADI P. DEEK
New Jersey Institute of Technology
rsb24@njit.edu
norbert.elliot@njit.edu
andrew.klobucar@njit.edu
Fadi.Deek@njit.edu

Assessment of educational outcomes through purchased tests is commonplace in the evaluation of individual student ability and of educational programs. Focusing on the assessment of writing performance in a longitudinal study of first-time, full-time students (n = 598), this research describes the design, use, and assessment of an open-source scoring platform. Augmenting usability testing, the research design relies on a framework of inter-reader agreement, inter-reader reliability, and coefficients of determination. The open-source, web-based portfolio assessment system yielded rates of agreement, reliability, and determination superior to the traditional paper-based portfolio assessment method. In addition, the system appears to be ideally suited to assess EPortfolios created to showcase student ability in digital environments: agreement range = 70% to 85%; reliability range = $\kappa = .67$ ($p < .01$) to $\kappa = .85$ ($p < .01$); coefficient of determination = $R^2 = .95$, $F(5, 34) = 118.59$ ($p < .01$). This novel and innovative application of an open source platform for outcomes assessment yields the foundation for a sound validity argument, the control of human error, and complete system transparency and flexibility. Future research directions point to the need for the design and assessment of an open-source system designed to capture complex constructs as they emerge in digital environments.

With its identification of "shortcomings of postsecondary institutions" in graduation rates, learning outcomes, and core literacy skills, the report of former Secretary of Education Margaret Spellings was an early signal that accountability has now become an enduring part of American higher education (U.S. Department of Education, 2006, p. 3). As a result of such criticism, commonly used measures of college-level general educational outcomes have been developed to provide information about student learning across institutional sites. Among these purchased tests are the ACT's Collegiate Assessment of Academic Proficiency, the Council for Aid to Education's Collegiate Learning Assessment, and the Educational Testing Service's Proficiency Profile. Based on these tests, sweeping conclusions about student performance have been drawn (Arum, Roksa, Kim, Potter, & Velez, 2011) and detailed critique of such conclusions has been offered (Astin, 2011; Haswell, 2012).

Such tempests are not new. As early as 1937, concerns regarding the efficacy of assessments were raised by Carl C. Brigham, creator of the SAT. "The pupil will gain if he is properly measured, but in the mad surge to measure two million pupils, no one is trying to describe just one pupil accurately" (p. 757). New is the role of open source platforms to provide a technological alternative to purchased tests, one that yields rapid assessment of student performance in both formative and summative settings within an evidence-based framework of validation. This paper describes a web-based portfolio assessment application that addresses the long-standing concerns articulated by Brigham by using state-of-the-art technology in the service of accurate assessment of the individual student. The case study focuses on the assessment of first-time, full-time student performance at a public science and technology research university. A complex construct to assess, writing ability provides a sufficiently challenging environment within which to test the capaciousness of open source development.

LITERATURE REVIEW

Establishing a framework for the present study demands attention to the nature of performance assessment, the role of digital environments in such assessment, and the demands of what are typically termed "next generation" assessments. Within an environment in which federal, regional, and state accountability demands are increasing, open source platforms such as the one developed and validated in this study may be positioned to yield substantial shareholder gains.

Position of Performance Assessment in Contemporary Measurement

Performance assessments, as Lane and Stone (2006) note, have been present in educational measurement since the middle of the twentieth century. Over the past decade attention to this unique assessment form has increased. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) define performance assessments in terms of construct coverage as attempts "to emulate the context or conditions in which the intended knowledge or skills are actually applied" (p. 137). Extending the concept to address extension of test results, Kane, Crooks, and Cohen (1999) emphasize the "close similarity between the type of performance that is actually observed and the type of performance that is of interest" (p. 388). The link between performance assessment and its intended use is characteristic of alignments that are part of the contemporary assessment milieu.

Within this environment, state-of-the-art evaluation defines validation as a process intended to provide evidence to support assessment use. Interpretative arguments follow the Toulmin (1958) model in which claims are made that are supported with information. Warrants—generalizations made to justify links between particular claims and the data that support them—are offered when explicit connections of proof are necessary. Qualifications are used in the model to express contingency (Mislevy, 2006, 2007). Accumulation of evidence to provide a sound scientific basis for proposed score interpretation is of increasing importance (Kane, 2006). Taken together, the robust nature of performance assessments and the insistence on validation of test use establish a rich framework for the present study.

Role of Digital Environments in Performance Assessment

It is interesting to observe that Lane and Stone (2006) address writing performance as an example of an application of performance assessment. Taught within the humanities, the prose essay continues to be the fundamental discursive form for undergraduate education, holding fast to both a cultural and pedagogical ethos that strongly distinguishes artistic modes of rhetorical discourse from other communication and media formats. Sustained by growing and increasingly diverse print markets from the mid-nineteenth to late twentieth century, prose writing itself has continued to evolve as an important cultural practice. Its social role as the primary signifier of modern individual reflection, self-sovereignty, and intellectual reasoning remained essentially unchallenged in the first half of the twentieth century.

At the same time, complementing the mid-twentieth century emergence and subsequent development of new communication and media technologies, these same discursive competencies, with their emphasis on expression and a single, authorial voice, have themselves become targets of significant cultural and politico-economic interrogation. Accordingly, while traditional views of rhetorical discourse continue to dominate composition and writing pedagogy, their capacity to direct newer communication and media practices, especially in the digital era, is increasingly less compelling. Innovative digital writing practices in the early twenty-first century define a significantly different communication environment, one in which newer multi-modal approaches to knowledge, gaming-based learning paradigms, and dynamic, network-informed understanding of communication supplant older, print-based measurements of literacy (Rice, 2007). Findings presented by the Inter/National Coalition for Electronic Portfolio Research (Cambridge, Cambridge, & Yancey, 2010) suggest that composing in electronic environments provides scaffolding and connection and thus invites the meaning-making characteristics of deep learning (Yancey, 2009, 2012). Especially relevant to the present study are efforts to introduce more digital media tools and assignments—blogging, podcasting and Wikis—into the curriculum and to use EPortfolios as part of electronic learning technologies that encourage students to consider document design, information organization, and social networking as increasingly integral to increased proficiency in writing performance (Klobucar, Deane, Elliot, Ramineni, Deess, & Rudniy, 2012). Assessment of such work using related digital technologies is conceptualized as a new form of educational measurement in which technology is neither a platform nor vehicle for the assessment but, instead, integral to the assessment design, use, and validation.

Anticipation of Next Generation Assessments

As defined by Jones and Vickers (2011), next generation assessments will be designed to capture the range of knowledge and skills students need for success in post-secondary education and 21st century careers. As such, the assessments will be increasingly robust and include more performance items and tasks. Performance tasks in particular are desirable, Jones and Vickers note, because students can demonstrate their ability to perform research, to apply knowledge through critical thinking skills, and to undertake analysis to solve problems; therefore, students can demonstrate their capability to complete authentic, real-world tasks. Because performance tasks provide a direct link between the act of learning and its assessment, technology developed in the service of digitally-based performance assessment is especially promising in its system management and scoring technology capabilities, as well as its implicitly distributed orientation allowing "anywhere" access to management functions, including validity, calibration, and reporting (p. 29). Focusing on the automated scoring capability of such systems, Bennett (2011) has stressed conceptualization as a continuum, "where the most innovative (and least trustworthy) methods are always paired with well-supervised human scoring and the least innovative (but most trustworthy) methods run with only human checking of quality-control samples" (p. 18). Examination of human-machine discrepancies, Bennett argues, will help identify weaknesses in both automated and human process, allowing developers and program managers to focus on quality improvement.

Critique of such systems has been vocal by the writing assessment community. Ericsson and Haswell (2006) have noted both the constraints of construct representation in the use of automated essay scoring and the lack of transparency in system development. Under conditions of timed writing, critics claim, only a very limited aspect of the construct—basic knowledge of conventions or limited organizational structure—can be assessed. Lack of transparency in next generation assessments can be understood by reference to Latour's description of "blackboxing": "When a machine runs efficiently, when a matter of fact is settled, one need focus only on its inputs and outputs and not on its internal complexity" (p. 204). If the framework for an automated writing assessment system is not transparent, for example, the system may default to a simple measure of word count and report that as proficiency, as Perelman (2012) has claimed of such systems. While it is not the purpose of this study to determine the resolution of such complex issues, open source platforms hold the potential to play a unique role in next generation assessments.

Role of Open Source in Next Generation Assessments

Open source software is increasingly providing the applications and systems infrastructure for the academic environment, and there are compelling reasons for dramatically expanding its use (Deek & McHugh, 2007). Similarly, open educational content and course development tools are serving the teaching and learning enterprise in a variety of ways. Open source materials that support the teaching of academic subjects are widely available on the web for use by instructors or self-learners. The advantages that can come from using open software/content in this budget-constrained era of education are obvious.

While public perception of the importance of open source is increasingly positive, there remains an inadequate awareness of the applicability of open source, the usability of these products and, more importantly, a concern regarding their cost benefits (Zhao, Deek, & McHugh, 2010; Rajanen & Iivari, 2010). However, in education, there is a strong predisposition for open products like student information and financial systems, course management systems, and portal frameworks, although until now, tools for facilitating assessment and measurement of student learning were still perceived to be not competitive with commercial software. WebPAA—a web-based portfolio assessment application—is a unique offering in this area, and it is readily available to other institutions interested in adapting the platform to suit their particular assessment requirements.

OPEN SOURCE SYSTEM DESIGN OF WEBPAA

The WebPAA system was designed and experimentally field-tested in the spring of 2010 and operationalized in the fall of 2010. The system followed a scenario-based design that distinguished between paper (traditional) and web-based (digital) models, allowing incorporation of the decision-basis for writing assessment into the latter while focusing on limiting known sources of error in the former. The WebPAA was able to be designed locally—an important advantage to open source platform use—to yield an efficient relational database that, if validated successfully, could be beneficial for future assessments similarly evidence-based in design.

Development was performed locally using the XAMPP package of tools—an open-source, cross-platform distribution containing the most common web development technologies in a single package including the Apache HTTP server, MySQL, PHP, and Perl. (Hence, the popular acronym of a cross or X, platform server is derived.) After localized testing, the completed application was uploaded to the cloud using Amazon Web Services (AWS) and the Amazon Elastic Compute Cloud (EC2). By locating the application in the cloud, there was no need for a dedicated server, thus making WebPAA a cost-effective option that provides not only dependability but also flexibility.

The development of WebPAA followed a scenario-based design technique (Carroll, 1999; Goodwin, 2009; Rosson & Carroll, 2002) that utilizes descriptive scenarios to aid in user interface and system design. Through the use of scenarios describing system usage at various levels of hierarchy to various stakeholders, scenario-based design enables the creation of a technology artifact that facilitates "new ways of doing things and new things to do" (Carroll, 1999). A critical consideration during development was the system's learnability and memorability (Holzinger, 2005). Learnability focuses on usability issues that allow a new user to quickly become comfortable with the system; memorability stresses the ease with which a user can return to the system after a prolonged period and be able to easily remember how to use the system. Because portfolio assessment is performed once per semester by professors who have many other responsibilities, both learnability and memorability were considered critical to the success of WebPAA.

To ensure the creation of an assessment technology that facilitated innovation and action, research scenarios included both efficacy and aesthetic variables (Collins, 2010). Efficacy variables included task completion, navigation, and textual descriptions; aesthetic variables included the effective use of typography, the overall layout of the pages, and the color and design used throughout the application. Interviews with the various stakeholders of the portfolio assessment process were conducted in the spring of 2010 with both raters (n = 5) and administrators (n = 4) to create scenarios of usage. Usability testing was performed simulating a real portfolio scoring environment; novice and experienced faculty members, as well as administrators, were invited to one of three usability testing sessions. All participants were located in a single room, each at a computer workstation. Sample student portfolios were provided, and raters were asked to rate several students while the administrators monitored their progress. Situations requiring adjudication of student work were simulated to ensure the usability of the adjudication process.

In examining administrator-level review of the scenario-based design, it became apparent that administrators encountered difficulties in using the portfolio assessment application. This is not surprising because the administrators have more tasks and menu options than the raters. Conversely, raters—required to perform only a single task, albeit a complex one—encountered fewer difficulties in using the system. Based upon these scenarios, rapid prototype models were created to solicit feedback from the stakeholders prior to actual system development (Dumas & Redish, 1999; Jones & Richey, 2000).

Based upon the scenarios and rapid prototype models, the development of the application was completed in a modular fashion with a focus on following open source software guidelines to enable present use by other institutions and future collaboration with other researchers. This developmental framework ensures that the application can be shared with other institutions which could adapt the instrument to their own programs. The application can be made available as a kernel in SourceForge.net so that others may use our research to develop assessments appropriate to their institutional requirements.

METHODOLOGY

Despite attention to the design of the digital system and its usability, the true test of the WebPAA system rested in its ability to allow reliability and model building of scores similar to that of the paper-based system. While scenario-based usability testing was a necessary precondition to validation of the WebPAA so that its future use could be determined, four rigorous tests were determined during the experimental stage of the system's use during the spring of 2010 and the operational examination of the system during the fall of 2010.

Research Questions

This study addresses the following research questions:

1. Will the WebPAA yield similar rates of inter-reader agreement to the traditional, paper-based scoring system?
2. Will the WebPAA yield similar rates of inter-reader reliability to the traditional, paper-based scoring system?
3. Will the WebPAA yield similar coefficients of determination to the traditional, paper-based scoring system?
4. If scored asynchronously with the WebPAA, will both hard copy and EPortfolios be scored at similar rates of inter-reader agreement, inter-reader reliability, and coefficients of determination to portfolios scores synchronously?

Inter-reader Agreement and Inter-reader Reliability Analysis

Classified by Stemler (2004) as a consensus estimate, inter-reader agreement is based on the assumption that trained readers should be able to come to established levels of agreement about how to apply various levels of a scoring rubric to the observed behaviors. If two judges come to exact agreement on how to use the rating scale to score behaviors, then the two judges may be said to share a common interpretation of the construct. In the present study, all levels of agreement were recorded. A variable on a portfolio would be assigned to a third reader if that reader disagreed with the initial rating by more than one point. That is, a score of 6 by a first reader and a score of 5 by a second reader was judged as in agreement; however, a score of 6 by a first reader and a score of 4 by a second reader—beyond adjacency—would be sent to a third reader.

Classified by Stemler (2004) as a consistency estimate, measures such as a Pearson product moment correlation and a weighted Kappa statistic indicate the degree to which a pattern of high and low scores is similar among raters (Brown, Glaswell, & Harland, 2004). In this study, both the Pearson correlation and the weighted Kappa were used (Abedi, 1996; Cohen, 1968; Fleiss & Cohen, 1973), with the weighted kappa somewhat privileged because of its sensitivity to differences in rater means and variances (Schuster, 2004).

While definitions are readily available for these measures, standards by which to judge their achievement in portfolio assessment are not. DiPardo, Storms, and Selland (2011) have identified adjacent rates of inter-reader agreement between 88% and 90% on a single variable within a multi-trait rubric as strong. In the technical documentation to the National Assessment of Educational Progress (NAEP) (2010), an agreement rate of 60% is considered an acceptable result when scoring on a complex six-point scale such as the one used in this study. Regarding interpretation of the Pearson product moment correlation on a portfolio containing only two scores, Nystrand, Cohen, and Dowling (1993) judge correlations of .38 as low, .66 as slightly higher, and .86 as substantially higher. Similar ambiguity is present for use of the weighted Kappa statistic. The NAEP (2010) notes that Kappa statistics should be higher than 0.6 for a six-point scale, but the application of that standard to portfolio scores of the kind described in this study are unknown.

In general, the strength of agreement range proposed by Landis and Koch (1977) provides a relative strength of agreement associated with the Kappa scale: < 0.00 = poor; 0.00 - 0.20 = slight; 0.21 - 0.40 = fair; 0.41 -

0.60 = moderate; 0.61 - 0.80 = substantial; and 0.81 – 1.00 = almost perfect. While these labels are helpful, for the purpose of determining future use of the WebPAA, the determining standard was that the digital system must meet or exceed the inter-reader reliability rates achieved in previous use—in this case, the fall of 2008, the spring of 2009, and the fall of 2009.

Coefficients of Determination Analysis

While reliability was the major concern of writing assessment researchers during the twentieth-century, validity has emerged as the chief concern during the twenty-first century. Within this framework, single scores of complex performances such as portfolios have drawn criticism. Murphy and Yancey (2010) have summarized objections in two areas: failure of a single score to reflect the writing construct (Williamson, 1993); and conflation of the complex variables associated with the writing construct that yield little diagnostic information (Hamp-Lyons, 1991). To these concerns may be added what Atkinson (2004) has termed a signaling effect, the message that assessment sends to students. As such, a multi-trait model suggests to students that writing is "a rich, multifaceted, meaning-making activity that occurs over time and in a social context, an activity that varies with purpose, situation, and audience and is improved by reflection on the written product and on the strategies used in creating it" (Camp, 1996, p. 135). As well, assessing writing in a digital framework sends the signaling effect that exploration of the variables of writing ability is welcome as these emerge in digital environments (Fraiberg, 2010).

In pursuit of a multi-trait variable model, during both the traditional period of the study (fall 2008, spring 2009, and fall 2009), and its experimental phase (spring 2010), the assessment design identified and used four independent variables: critical thinking; revising and editing; content and organization; and sentence construction and mechanics. As in the case with previous research on first year students (Elliot, Briller, & Joshi, 2007; Elliot, Deess, Rudniy, & Joshi, 2012), these four variables were taken as independent (predictor, or *X*) variables that contributed to the holistic (outcome, or *Y*) score. As such, each portfolio received five scores based on a Likert scale ranging from 6 (the highest score) to 1 (the lowest score). However, during the fall 2010 operational stage of the study, the assessment evolved in order to reflect national standards of best practice associated with the Outcomes Statement adopted by the Council of Writing Program Administrators (2004, 2008). During this period, paper portfolios used rhetorical knowledge, critical thinking, writing processes, and knowledge of conventions as the independent variables associated with a holistic score. During this phase a fifth independent variable—composting in electronic environments—was added to capture experiences of writing associated with digital portfolios.

Synchronous and Asynchronous Scoring

During the fall of 2010, an additional scoring dimension was added to the assessment framework: asynchronous scoring. Just as writing in digital environments is a new topic for theoretical and practical experiment (Rice, 2007), so too is assessing writing in these environments (Neal, 2011). As well, research in the teaching of writing in distributed environments holds the potential to identify methods by which the digital divide may be bridged and a more fully integrated vision of lifelong learning may occur across time and circumstance (Neff & Whithaus, 2008). When writing is assessed within networked technological systems, raters who are trained to score in online distributed environments may be less likely to exhibit centrality (compression of ratings toward the center of the scoring distribution) and inaccuracy (low rater consistency) effects (Wolfe & McVay, 2010). As such, along with inter-reader agreement, inter-reader reliability, and coefficient of determination analysis, comparison of synchronous and asynchronous scoring was identified as a fourth analysis to determine the potential use of the system.

Sampling Plan Design: Spring 2010 and Fall 2010

Fall 2008, spring 2009, fall 2009, and spring 2010 portfolio scores were obtained by a sampling plan based on a 95% confidence interval of all admitted students (Elliot, Briller, Johsi, 2008; Johnson & Elliot, 2010). So, for the fall of 2008, when 939 full-time, first-time (FTFT) students were admitted, the sampling plan required that 181 portfolios would have to be read if we were to be confident that our scores were representative of the admitted class. In the fall of 2009, 1021 FTFT students were admitted; of that total, 151 portfolios were read. During fall 2008, spring 2009, fall 2009, and spring 2010, all portfolios were read twice.

While sampling plan design did not change for the fall of 2010 when 1006 students were admitted, the human cost of reading each portfolio in the sampling plan twice became excessive. Following the recommendation of Gay, Mills, and Airasian (2011) to use sample sizes of at least 30 for correlation research, we randomly selected 44 portfolios to be read twice for the fall of 2010. However, since we had no experience reading EPortfolios during that semester, each of the 44 was read twice.

RESULTS

Three types of evidence are provided to justify the continued use of the WebPAA system. In each case, the open source system yielded results of similar or superior quality in inter-reader agreement, inter-reader reliability, and coefficient of determination.

Inter-reader Agreement Evidence

As shown in Table 1, the WebPAA system allowed inter-reader rates of agreement similar to those of the traditional system. Under the traditional system, the lowest rate of inter-reader agreement reached in exact and adjacent agreement—those portfolios requiring no adjudication—ranged from a low of 79% for revising and editing in the fall of 2009 to a high of 96% for sentence construction and agreement in the fall of 2008. During the experimental phase of the WebPAA in the spring of 2010, agreement ranged from a low of 72% for critical thinking to a high of 92% for writing processes and the holistic score. During the operational phase of the WebPAA in the fall of 2010, agreement ranged from a high of 98% for the holistic score on the paper-based portfolios to a low of 70% for the EPortfolio variable of composing in electronic environments. Setting that new variable aside, the lowest rate of agreement—75% on rhetorical knowledge on EPortfolios—was comparable to that of the traditional system. Indeed, on the paper-based portfolios, the WebPAA rates of agreement exceeded those of the traditional system, with a high of 98% for the holistic score and a low of 82% for writing processes. Such rates of agreement meet and exceed the NAEP (2010) agreement rate of 60%.

Table 1
Descriptive Statistics and Inter-rater Agreement Indicators, Fall 2008 to Fall 2010 (n = 598)

Descriptive statistics				Inter-rater agreement indicators			
Competency	Range	Mean	*SD*	Exact agreement	Exact plus adjacent	Scores differ by 2	Scores differ by 3
Fall 2008 Traditional Print Scoring (n = 181)							
1. Critical Thinking	4, 12	8.12	1.51	86 (48%)	87 (48%)	8 (4%)	0 (0%)
2. Revising and Editing	2, 12	7.41	1.77	76 (42%)	77 (43%)	27 (14%)	1 (1%)
3. Content and Organization	3, 12	8.06	1.54	88 (49%)	83 (46%)	10 (5%)	0 (0%)
4. Sentence Construction and Mechanics	3, 12	7.77	1.7	72 (40%)	101 (56%)	8 (4%)	0 (0%)
5. Holistic Score	3, 12	8.04	1.64	90 (49%)	79 (44%)	12 (7%)	0 (0%)
Spring 2009 Traditional Print Scoring (n = 103)							
1. Critical Thinking	2, 11	7.84	1.74	60 (58%)	38 (37%)	5 (5%)	0 (0%)
2. Revising and Editing	2, 11	6.94	2.11	42 (41%)	47 (45%)	10 (10%)	4 (4%)
3. Content and Organization	3, 11	7.86	1.6	56 (54%)	42 (41%)	5 (5%)	0 (0%)
4. Sentence Construction and Mechanics	3, 11	7.9	1.52	54 (52%)	37 (36%)	12 (12%)	0 (0%)
5. Holistic Score	2, 11	7.82	1.8	53 (50%)	45 (45%)	5 (5%)	0 (0%)
Fall 2009 Traditional Print Scoring (n = 151)							
1. Critical Thinking	3, 12	7.6	1.84	61 (40%)	74 (49%)	15 (10%)	1 (1%)
2. Revising and Editing	2, 11	6.35	2.32	56 (37%)	63 (42%)	31 (20%)	1 (1%)
3. Content and Organization	2, 12	7.44	1.9	69 (46%)	68 (46%)	13 (7%)	1 (1%)
4. Sentence Construction and Mechanics	2, 11	7.72	1.64	75 (51%)	63 (42%)	13 (7%)	0 (0%)
5. Holistic Score	2, 12	7.47	1.84	72 (48%)	65 (43%)	14 (9%)	0 (0%)

Table 1 Continued

Descriptive statistics				Inter-rater agreement indicators			
Competency	Range	Mean	*SD*	Exact agreement	Exact plus adjacent	Scores differ by 2	Scores differ by 3
Spring 2010 Experimental WebPAA Synchronous Paper Portfolios (n = 79)							
1. Rhetorical Knowledge	3, 11	7.87	1.46	29 (37%)	35 (46%)	13 (16%)	1 (1%)
2. Critical Thinking	2, 10	6.53	2.27	21 (27%)	36 (45%)	19 (24%)	3 (4%)
3. Writing Processes	3, 11	7.89	1.48	32 (40%)	41 (52%)	6 (8%)	0 (0%)
4.Conventions	3, 11	8.11	1.45	32 (40%)	39 (50%)	7 (9%)	1 (1%)
5. Holistic Score	2, 11	7.49	1.62	31 (40%)	41 (52%)	5 (6%)	2 (2%)
Fall 2010 Operational WebPAA Asynchronous Paper Portfolios (n = 44)							
1. Rhetorical Knowledge	6, 10	8.41	1.12	30 (68%)	9 (21%)	5 (11%)	0 (0%)
2. Critical Thinking	5, 11	8.3	1.27	21 (48%)	19 (43%)	3 (7%)	1 (2%)
3. Writing Processes	3, 11	7.39	1.87	16 (36%)	20 (46%)	8 (18%)	0 (0%)
4.Conventions	4, 11	8.0	1.55	22 (50%)	18 (41%)	4 (9%)	0 (0%)
5. Holistic Score	5, 12	8.14	1.42	20 (46%)	23 (52%)	4 (2%)	0 (0%)
Fall 2010 Operational WebPAA Asynchronous EPortfolios (n = 40)							
1. Rhetorical Knowledge	2, 11	7.48	2.49	18 (45%)	12 (30%)	10 (25%)	0 (0%)
2. Critical Thinking	2, 12	7.05	2.48	22 (55%)	10 (25%)	8 (20%)	0 (0%)
3. Writing Processes	2, 12	7.17	2.39	23 (58%)	11 (27%)	6 (15%)	0 (0%)
4.Conventions	2, 11	8.05	2.18	21 (53%)	13 (32%)	6 (15%)	0 (0%)
5. Composing in Electronic Environments	2, 11	5.95	2.93	17 (43%)	11 (27%)	10 (25%)	2 (5%)
6. Holistic Portfolio Score	2, 12	7.23	2.82	17 (43%)	15 (37%)	5 (13%)	3 (7%)

Inter-reader Reliability Evidence

As shown in Table 2, inferential statistics from the Pearson product moment correlation and the weighted kappa statistic yielded similar trends to the inter-reader agreement analysis. Under the traditional system, the lowest rate of non-adjudicated inter-reader reliability measured by the Pearson product moment correlation (two tailed), $r = .41$ ($p < .01$), occurred in revising and editing for the fall of 2008. The highest rate, $r = .67$ ($p < .01$), occurred for the holistic score for the spring of 2009. Under conditions of adjudication, the lowest rate, $r = .62$ ($p < .01$), occurred in critical thinking in the fall of 2008. The highest rate of inter-reader reliability, $r = .82$ ~~(π~~ < .01), occurred for revising and editing in the fall of 2009. During the experimental phase of the WebPAA in the spring of 2010, non-adjudicated inter-reader reliability ranged from a correlation of $r = .27$ ($p < .01$) on rhetorical knowledge to a high of $r = .50$ ($p < .01$) on the holistic score. Adjudicated scores during this phase ranged from $r = .56$ ($p < .01$) on rhetorical knowledge to $r = .75$ ($p < .01$) on critical thinking. During the operational phase of the WebPAA in the fall of 2010, the non-adjudicated critical thinking variable on the paper portfolios did not achieve statistical significance, the sole occurrence during the investigation. The highest correlation for the non-adjudicated holistic score, $r = .59$ ($p < .01$), was comparable to correlations on that variable for the fall of 2008 and the fall of 2009. Notable were the higher ranges of correlations—the highest in the study—for the EPortfolios, with a low of $r = .60$ ($p < .01$) on rhetorical knowledge to a high of $r = .71$ ($p < .01$) on critical thinking and the holistic score. Regarding the adjudicated scores, the ranges on the WebPAA are comparable to those of the traditional system, with a low of $r = .48$ ($p < .01$) on the critical thinking score for the paper portfolios to a high of $r = .84$ ($p < .01$) on the EPortfolio holistic score. The range of adjudicated scores as measured by the Pearson product moment correlation is notably high,

Inferential statistics from the weighted kappa yielded ranges similar to the Pearson product moment correlations. Under the traditional system, the lowest rate, $\kappa = .28$ ($p < .01$), occurred in revising and editing in the fall of 2008. The highest rate of non-adjudicated inter-reader reliability, $\kappa = .58$, $p < .01$, occurred for the critical thinking and holistic scores for the fall of 2009. Under conditions of adjudication, the lowest rate for the traditional system, $\kappa = .40$, $p < .01$, occurred in sentence construction and mechanics during the fall of 2008. The highest inter-reader reliability rate, $\kappa = .82$, $p < .01$, occurred in revising and editing in the fall of 2009. During the experimental phase of the WebPAA in the spring of 2010, non-adjudicated weighted kappa statistics ranged from a low of $\kappa = .27$ ($p < .01$), on rhetorical knowledge to a high of $\kappa = .50$ ($p < .01$), on the holistic score.

Table 2
Inter-rater Reliability Indicators, Fall 2008 to Fall 2010
(n = 598)

Competency	Non Adjudicated Pearson	Adjudicated Pearson	Non-Adjudicated Weighted Kappa	Adjudicated Weighted Kappa
Fall 2008 Traditional Print Scoring (n = 181)				
1. Critical Thinking	.52**	.62**	.34**	.41**
2. Revising and Editing	.41**	.71**	.28**	.49**
3. Content and Organization	.54**	.65**	.33**	.41**
4. Sentence Construction and Mechanics	.62**	.66**	.36**	.40**
5. Holistic Score	.58**	.67**	.41**	.46**
Spring 2009 Traditional Print Scoring (n = 103)				
1. Critical Thinking	.65**	.76**	.52**	.58**
2. Revising and Editing	.51**	.78**	.37**	.56**
3. Content and Organization	.62**	.72**	.45**	.51**
4. Sentence Construction and Mechanics	.46**	.66**	.45**	.51**
5. Holistic Score	.67**	.76**	.47**	.53**
Fall 2009 Traditional Print Scoring (n = 151)				
1. Critical Thinking	.50**	.71**	.58**	.71**
2. Revising and Editing	.51**	.82**	.50**	.82**
3. Content and Organization	.54**	.71**	.54**	.70**
4. Sentence Construction and Mechanics	.53**	.68**	.53**	.68**
5. Holistic Score	.58**	.71**	.58**	.71**

Note: all t-tests are two-tailed
* $p < .05$, ** $p < .01$

Table 2 Continued

Competency	Non Adjudicated Pearson	Adjudicated Pearson	Non-Adjudicated Weighted Kappa	Adjudicated Weighted Kappa
Spring 2010 Experimental WebPAA Synchronous Paper Portfolios (n = 79)				
1. Rhetorical Knowledge	.27**	.56**	.27**	.55**
2. Critical Thinking	.44**	.75**	.43**	.74**
3. Writing Processes	.41**	.60**	.41**	.60**
4.Conventions	.36**	.57**	.36**	.57**
5. Holistic Score	.50**	.65**	.50**	.64**
Fall 2010 Operational WebPAA Asynchronous Paper Portfolios (n = 44)				
1. Rhetorical Knowledge	.38**	.64**	.38**	.63**
2. Critical Thinking	.29 *ns*	.48**	.29*	.47**
3. Writing Processes	.47**	.72**	.46**	.70**
4.Conventions	.53**	.64**	.5**	.63**
5. Holistic Score	.59**	.64**	.58**	.62**
Fall 2010 Operational WebPAA Asynchronous EPortfolios (n = 40)				
1. Rhetorical Knowledge	.60**	.71**	.53**	.67**
2. Critical Thinking	.71**	.86**	.69**	.85**
3. Writing Processes	.68**	.83**	.68**	.83**
4.Conventions	.66**	.80**	.61**	.79**
5. Composing in Electronic Environments	.66**	.77**	.62**	.75**
6. Holistic Portfolio Score	.71**	.84**	.66**	.81**

Note: All t-tests are two-tailed
* p < .05, ** p < .01

Adjudicated statistics ranged from a low of $\kappa = .55$ ($p < .01$) on rhetorical knowledge to a high of $\kappa = .74$ ($p < .01$) on critical thinking. During the operational phase of the WebPAA in the fall of 2010, the lowest non-adjudicated weighted kappa statistic, $\kappa = .29$ ($p < .05$), achieved statistical significance, while the highest, $\kappa = .69$ ($p < .01$) on critical thinking for the EPortfolios, achieved the highest statistic in the study for any variable. In a similar fashion, both the lowest and the highest weighted kappa statistics for adjudicated scores—for critical thinking ($\kappa = .47$, $p < .01$) on the paper portfolios and for critical thinking on the EPortfolios ($\kappa = .85$, $p < .01$)—achieved ranges similar to those of traditional scoring.

Under adjudicated conditions, the NAEP (2010) specifications recommend that Kappa statistics should be higher than 0.6 for a six-point scale. This level of Kappa statistics is approximated and, in the majority of cases, exceeded in both the experimental and operational stages of the present study. Remarkably, even the non-adjudicated scores for holistic scores, the outcome variable for the study, approximate or exceed this standard. Interpreting the scores in the relative strength of agreement framework offered by Landis and Koch (1977), the majority of the adjudicated scores are moderate.

Coefficients of Determination Evidence

As shown in Table 3, coefficients of determination yielded similar trends to the agreement and reliability analyses. Under the traditional system, each of the independent variables predicted the holistic score at statistically significant levels, with a range of $\beta = .14$ ($p < .01$) in revising and editing for each traditional print scoring to a high of $\beta = .5$ ($p < .01$) for critical thinking in the spring of 2009. Coefficients of determination were high, ranging from a model that predicted 75% of the variance in fall of 2008 ($R^2 = .75$, $F(4, 176) = 134.36$, $p < .01$) to a prediction of 84% in the fall of 2009 ($R^2 = .84$, $F(4, 146) = 187.59$, $p < .01$). During the experimental phase of the WebPAA in the spring of 2010, the prediction rate remained high ($R^2 = .78$, $F(4, 74) = 65.75$, $p < .01$). However, neither the critical thinking nor the conventions variable achieved statistical significance, thus failing to contribute to model prediction. A similar pattern remained during the operational phase of the WebPAA in the fall of 2010. In the paper portfolios scored asynchronously, the critical thinking independent variable did not achieve statistical significance in predicting the holistic score. In the EPortfolios scored asynchronously, the writing processes independent variable did not achieve statistical significance in predicting the holistic score. Nevertheless, the WebPAA yielded similar coefficients of determination to the traditional system on both paper portfolios (R^2 = .8[illegible], F(4, 39) = 38.66, p < .01) and EPortfolios (R^2 = .95, F(5, 34) = 118.59, p < .01). Significantly, coefficients of determination increased in the experimental and operational phase of the study as readers became more familiar with the system.

Table 3

Regression Indicators, Fall 2008 to Fall 2010

(n = 598)

Model	Standardized beta coefficient (β) for predictor variables (X)	Multiple correlation coefficient (R) for model ($X \rightarrow Y$)	Coefficient of determination (R^2) for model ($X \rightarrow Y$)	Fisher's F ratio (F) for model ($X \rightarrow Y$)
Fall 2008 Traditional Print Scoring (n = 181)				
1. Critical Thinking (X_i)	.31 **			
2. Revising and Editing (X_{ii})	.14 **			
3. Content and Organization (X_{iii})	.33 **			
4. Sentence Construction and Mechanics (X_{iv})	.21 **			
5. Holistic Score (Y)		R = .87 **	R^2 = .75 **	F = (4, 176) 134.36 **
Spring 2009 Traditional Print Scoring (n = 103)				
1. Critical Thinking (X_i)	.5 **			
2. Revising and Editing (X_{ii})	.14 **			
3. Content and Organization (X_{iii})	.20 *			
4. Sentence Construction and Mechanics (X_{iv})	.16 *			
5. Holistic Score (Y)		R = .90 **	R^2 = .81 **	F = (4, 98) 102.62 **

ns = not statistically significant

* p < .05, ** p < .01

Table 3 Continued

Model	Standardized beta coefficient (β) for predictor variables (*X*)	Multiple correlation coefficient (*R*) for model ($X \rightarrow Y$)	Coefficient of determination (R^2) for model ($X \rightarrow Y$)	Fisher's *F* ratio (*F*) for model ($X \rightarrow Y$)
Fall 2009 Traditional Print Scoring (n = 151)				
1. Critical Thinking (X_{i})	.39**			
2. Revising and Editing (X_{ii})	.14**			
3. Content and Organization (X_{iii})	.33**			
4. Sentence Construction and Mechanics (X_{iv})	.18**			
5. Holistic Score (*Y*)		*R* = .92 **	R^2 = .84 **	*F* = (4, 146) 187.59 **
Spring 2010 Experimental WebPAA Synchronous Paper Portfolios (n = 79)				
1. Rhetorical Knowledge (X_{i})	.23**			
2. Critical Thinking (X_{ii})	.11 *ns*			
3. Writing Processes (X_{iii})	.57**			
4.Conventions (X_{iv})	.10 *ns*			
5. Holistic Score (*Y*)		*R* = .88 **	R^2 = .78 **	*F* = (4, 74) 65.75 **
Fall 2010 Operational WebPAA Asynchronous Paper Portfolios (n = 44)				
1. Rhetorical Knowledge (X_{i})	.61**			
2. Critical Thinking (X_{ii})	.06 *ns*			
3. Writing Processes (X_{iii})	.20 *			
4.Conventions (X_{iv})	.20 *			
5. Holistic Score (*Y*)		*R* = .89 **	R^2 = .80 **	*F* = (4, 39) 38.66 **

ns = not statistically significant

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

Table 3 Continued

Model	Standardized beta coefficient (β) for predictor variables (X)	Multiple correlation coefficient (R) for model ($X \rightarrow Y$)	Coefficient of determination (R^2) for model ($X \rightarrow Y$)	Fisher's F ratio (F) for model ($X \rightarrow Y$)
Fall 2010 Operational WebPAA Asynchronous EPortfolios (n = 40)				
1. Rhetorical Knowledge (X_i)	.33**			
2. Critical Thinking (X_{ii})	.24*			
3. Writing Processes (X_{iii})	-.02 *ns*			
4.Conventions (X_{iv})	.23**			
5. Composing in Electronic Environments	.23**			
6. Holistic Score (Y)		R = .97 **	R^2 = .95 **	F = (5, 34) 118.59 **

ns = not statistically significant
* p < .05, ** p < .01

Comparative Evidence of Synchronous and Asynchronous Scoring

Setting aside the independent variable of composing in electronic environments, the three previous analyses reveal that portfolios scored asynchronously during the operational phase of the WebPAA in the fall of 2010 yielded higher ranges of inter-reader agreement than portfolios scored synchronously. Ranges of inter-reader reliability measured by the Pearson product moment coefficient on paper portfolios scored synchronously suggests that initial non-adjudicated readings were more aligned in the traditional system; adjudicated scores suggested similar correlations. Ranges in inter-reader reliability measured by the weighted Kappa coefficient revealed similar ranges under both non-adjudication and adjudication. However, ranges of inter-reader reliability on the EPortfolios read with the WebPAA achieved the consistently highest set of ranges on the weighted Kappa statistic, under both non-adjudicated and adjudicated conditions, in the study. While coefficients of determination revealed similar model predictions under both systems, EPortfolios read with the WebPAA achieved the highest rate of prediction—95%—in the study.

CONCLUSIONS, LIMITATIONS, AND SUGGESTIONS FOR FURTHER RESEARCH

The objective of the WebPAA system was to develop a robust platform that enables the portfolio assessment process to be implemented through web-based technology. Focusing on learnability and memorability usability features, the system was designed to ensure that raters could quickly and easily begin using the application whenever a portfolio rating session was initiated. Additionally, the system was developed to allow both synchronous and asynchronous rating of portfolios, allowing raters the flexibility of completing their ratings at a time and location convenient for them.

As such, the study design yielded desired interpretative arguments (Mislevy, 2006, 2007; Toulmin, 1958) offered to shareholders who support the system (administrators), use it (instructors), and are impacted by its use (students). Adhering to contemporary validation practices, Tables 1, 2, and 3 serve as warrants to justify the following claims: the WebPAA yields similar rates of inter-reader agreement to the traditional, paper-based scoring system; the WebPAA yields similar rates of inter-reader reliability to the traditional, paper-based scoring system; the WebPAA yields similar coefficients of determination to the traditional, paper-based scoring system; and both hard copy and EPortfolios scored asynchronously yield similar rates of inter-reader agreement, inter-reader reliability, and coefficients' of determination to portfolios scores synchronously. As further evidence offered to validate extended use of the WebPAA, the system eradicated human error associated with paper-based scoring and database entering of those scores. The platform's potential for real-time score monitoring also holds the potential to allow control of centrality and inaccuracy, two important known sources of reader error identified by Wolfe and McVay (2010). Equally significant, the system afforded complete transparency through its design with the XAMPP package of tools and its ability to be publically deployed through SourceForge.net.

As expressions of contingency, qualifications are present along with the claims. As an *in situ* field experiment, the research captures a complex situational environment in which the variables of the construct were altered during the operational phase of the study. While conditions for assessment of the construct were not kept consistent, it is nevertheless apparent that the platform's capaciousness allowed it to yield information to support the validity argument. Second, because of complexity of the variable model and performance assessment conditions, inter-reader agreement and inter-reader reliability for non-adjudicated scores often fall below rates established in NAEP and by Landis and Koch (1977). Experiments of the sort recorded here must be continued before a standard can be reached by which to judge these forms of agreement and reliability of complex constructs such as writing ability. Third, because technology is integrated within the assessment framework, the research is an evaluation both of reader variance (human) and machine capability (technology). To tease out known sources of reader variance such as centrality and inaccuracy, further research as that of Wolfe and McVay (2010) using applications of Rasch modeling (Andrich, 1978) need to be undertaken. Once identified, the WebPAA could easily control for such known sources of error by real-time monitoring of the assessment. Finally, because an evidence-centered design framework has been used in this performance assessment, a validation process is described, not a final stamp of approval. Indeed, caution should be expressed that described in this research is an entire validation system, not merely advocacy for a platform that will solve all ills. Individuals wishing to validate the WebPAA for use in their own specific sites must be willing to perform studies similar to the one described here before their deployment of the system. Such, of course, is the nature of institutional-based research.

The final qualification of the present study demonstrates the need for a re-orientation to the assessment of student learning that encompasses both ideographic representation and nomothetic span (Embretson, 1993; Borsboom, 2005). While the research here describes validation undertaken at a specific institutional site with a defined construct, future research with open source systems such as WebPAA should focus on ways to share system development to span known and new constructs in a collaborative fashion. In the fields of Information Systems (IS) and Information Technology (IT), researchers are calling for increased utilization of an IT artifact when conducting research (Orlikowski & Iacono, 2001). WebPAA provides such an IT artifact which can be assessed, tested, and used as a scaffold for additional research in developing technology to support summative student assessment through measures other than multiple choice tests and short answer examinations.

Brigham's 1937 lamentation regarding modernist systems designed to measure entire populations yet failing to accurately describe just the singular pupil are addressed in the kinds of systems—both in design of construct representation and in technological systems—described in the present research. Open source platforms are ideally suited to serve the new era of performance assessment described by Jones and Vickers (2011) and Bennett (2011), while reckoning with the claims of the constrained construct representation and lack of transparency in system development levied by Ericsson and Haswell (2006). Indeed, in and of itself, the deeply contextualized nature of the assessment described in this research distinguishes it from purchased assessments emerging from federal and regional demands.

While there are many areas to examine, it is indeed possible that such assessments may yield a criterion variable of student performance that is equal, if not superior, to the traditional course grade model. As Willingham, Pollack, and Lewis (2002) have shown, disjuncture between student performance under assessment conditions and instructor assigned final grades is due to variables such as student engagement. In the complex socio-cognitive environment (Bandura, 2006) in which subjects such as writing are taught within humanistic disciplines (Flower, 1986), new methods of educational measurement will need to be developed if students are to learn complex skills—such as composing in digital environments—that will be needed for personal accomplishment and professional success. Lest we gloss over the results in Table 1, it is disheartening to document that the lowest scores recorded in the present study (*M* = 5.95, *SD* = 2.93) were those of students using EPortfolios in the operational phase of the study who were unable to demonstrate that they could compose in electronic environments. Trapped in a print-based world of the twentieth-century, without intervention these students will be unable to master the complex skill sets necessary to operate in the rich communicative environment of the twenty-first century. The framework of open source development, both in instruction and assessment, holds the potential to address such shortcomings. Required for further research is a structured-case framework for theory-building, a socio-cognitive view of learning, a vision for the integrative potential of educational measurement and open source methodology, and an articulated research model. We shall devote future work to questions left unanswered by this one.

References

Abedi, J. (1996). The interrater/test reliability system. *Multivariate Behavioral Research, 31*, 409–417.

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

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Arum, R., Roksa, J., Kim, J., Potter, D., & Velez, M. (2011). *Academic adrift: Limited learning on college campuses*. Chicago, IL: University of Chicago Press.

Astin, A. W. (2011, February 14). In "Academically Adrift," data don't back up sweeping claim. *Chronicle of Higher Education*. Retrieved from http://chronicle.com/

Atkinson, R. (2004, April 14). AERA public service award address. San Diego, CA: American Educational Research Association.

Bandura, A. (2006). Toward a psychology of human agency. *Perspectives on Psychological Science*, *1*, 164 -180.

Bennett, R.E. (2011). *Automated scoring of constructed-response literacy and mathematics items*. Washington, DC: Arabella Philanthropic Investment Advisors. Retrieved from http://www.ets.org/research/

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge, England: Cambridge University Press.

Brigham, Carl C. (1937). The place of research in a testing organization. *School and Society 11*, 756.

Brown, G.T., Glasswell, K., & Harland, D. (2004). Establishing reliability and validity of a national assessment tool for writing. *Assessing Writing*, 9, 105-121.

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

Camp, R. (1996). New views of measurement and new models for writing assessment. In E. M. White, W. D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 135–147). New York, NY: Modern Language Association.

Carroll, J. M. (1999). Five reasons for scenario-based design. *Proceedings of the 32nd Hawaii International Conference on System Sciences*. Retrieved from http://dl.acm.org

Collins, R. (2010). *Web-based portfolio assessment: Platform design for writing assessment* (Master's thesis). Retrieved from http://library.njit.edu/etd/

Cohen, J. (1968). Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213–220.

Council of Writing Program Administrators (2004. 2008). WPA outcomes statement. Retrieved from http://wpacouncil.org

Deek, F. & McHugh, J. (2007). *Open Source: Technology and policy*. Cambridge, England: Cambridge University Press.

DiPardo, A., Storms, B. A., & Selland, M. (2011). Seeing voices: Assessing writerly stance in the NWP Analytic Writing Consortium. *Assessing Writing*, *16*, 170-188.

Dumas, J., & Redish, J.C. (1999). *A practical guide to usability testing*. Portland, OR: Intellect Books.

Elliot, N., ~~V.~~ Briller, ~~and~~ ~~K.~~ Joshi. (2007). Analytic portfolio assessment: A program development model. *Journal of Writing Assessment 3*, 5-30.

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

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179-97.

Ericsson, P. F., & Haswell, R. (2006). *Machine scoring of student essays: Truth and consequences*. Logan, UT: Utah State University Press.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*, 613-619.

Flower, L. (1986). *The construction of negotiated meaning: A social cognitive theory of writing*. Carbondale and Edwardsville, IL: Southern Illinois University Press.

Fraiberg, S. (2010). Composition 2.0: Toward a multilingual, multimodal framework. *College Composition and Composition*, *62*, 100-126.

Gay, L. R., Mills, G. E., & Airasian, P. W. (2011). Educational research: Competencies for analysis and application. 10th ed. Reading, MA: Addison Wesley.

Goodwin, K. (2009). *Designing for the digital age: How to create human-centered products and services*. Indianapolis, IN: Wiley Publishing.

Hamp-Lyons, L. (1991). Scoring procedures for ESL context. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241-276). Norwood, NJ: Ablex.

Haswell, R. (2012). Methodologically adrift. [Review of the book *Academically adrift: Limited learning on college campuses*, by R. Arum & J. Roska.] *College Composition and Communication*, *63*, 487-491.

Holzinger, A. (2005). Usability engineering methods for software developers. *Communications of the ACM*, *48*, 71-74.

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

Jones, M., & Vickers, D. (2011). Considerations for performance scoring when designing and developing next generation assessments. White paper. Upper Saddle River, NJ: Pearson. Retrieved from http://www.pearsonassessments.com

Jones, T. S., & Richey, R. (2000). Rapid prototyping methodology in action: A developmental study. *Educational Technology Research and Development*, *48*, 68-30.

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

Kane, M., Crooks. T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*, 5-17.

Klobucar, A., Deane, P., Elliot, N., Ramineni, C., Deess, P., & Rudniy, A. (2012). Automated essay scoring and the search for valid writing assessment. In C. Bazerman, C. Dean, J. Early, K. Lunsford, S. Null, P. Rogers, & A. Stansell (Eds.), International advances in writing research.

Cultures, places, measures (pp. 103-119). Fort Collins, Colorado: WAC Clearinghouse/Anderson, SC: Parlor Press. Retrieved from http://wac.colostate.edu/books/wrab2011/chapter6.pdf

Landis J. R., & Koch G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.

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

Latour, B. (1999). *Pandora's hope: Essays on the reality of science studies*. Cambridge, England: Cambridge University Press.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257-305). Westport, CT: American Council on Education/Praeger.

Mislevy, R. (2007). Validity by design. *Educational Researcher 36, 8*, 463-69.

Murphy, S., & Yancey, K. B. (2008). Construct and consequence: Validity in writing assessment. In C. Bazerman (Ed.), *Handbook of research on writing: History, society, school, individual, text* (pp. 365-385). New York, NY: Lawrence Erlbaum.

National Assessment of Educational Progress (2010). Constructed-response interrater reliability. Retrieved from http://nces.ed.gov/nationsreportcard/tdw/analysis/initial_itemscore.asp

Neal, M. R. (2011). *Writing assessment and the revolution in digital technologies*. New York, NY: Teachers College Press.

Neff, J. M., & Whithaus, C. (2008). *Writing across distances and disciplines: Research and pedagogy in distributed learning*. New York, NY: Lawrence Erlbaum.

Nystrand, M., Cohen, A., & Dowling, N. (1993). Addressing reliability problems in the portfolio assessment of college writing. *Educational Assessment*, *1*, 53-70.

Orlikowski, W., & Iacono, C. S. (2001). Research commentary: Desperately seeking the "IT" in IT Research—A call to theorizing the IT artifact. *Information Systems Research*, *12*, 121-34.

Perelman, L. (2012). Mass-market writing assessment as bullshit. In N. Elliot & L. Perelman (Eds.), *Writing assessment in the twenty-first century: Essays in honor of Edward M. White* (pp. 425-437). New York, NY: Hampton Press.

Rajanen, M., & Iivari, N. (2010). Traditional usability costs and benefits: Fitting them into open source software development. *ECIS 2010 Proceedings*. Paper 154. Retrieved from http://aisel.aisnet.org

Rice, J. (2007). *The rhetoric of cool: Composition studies and the new media*. Carbondale, IL: Southern Illinois University Press.

Rosson, M. B., & Carroll, J. M. (2002). Scenario-based design. In J. Jacko & A. Sears (Eds.) *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications* (pp, 1032-1050). Mahwah, NJ: Lawrence Erlbaum Associates.

Schuster, C. (2004). A note on the interpretation of weighed kappa and its relations to other rater agreement statistics for metric scales. *Educational and Psychological Measurement*, *64*, 243-253.

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

Toulmin, S. E. (1958). *The uses of argument*. Cambridge, England: Cambridge University Press

U.S. Department of Education (2006) *A test of leadership: Charting the future of U.S. higher education.* Washington, DC: U.S. Department of Education. Retrieved from http://www2.ed.gov/

Yancey, K. B. (2009). Electronic portfolios a decade into the twenty-first century: What we know, what we need to know. *Peer Review*, *11*, 28-33.

Yancey, K. B. (2012). The rhetorical situation of writing assessment: Exigence, location, and the meaning of knowledge. In N. Elliot & L. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Edward M. White* (pp. 475-492). New York, NY: Hampton Press,

Williamson, M. (1993). An introduction to holistic scoring. In M. Williamson & B Huot (Eds.), *Validating holistic scoring of writing assessment. Theoretical and empirical foundations* (pp. 206-232). Creskill, NJ: Hampton.

Willingham, W. W., Pollack, J. M., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement*, *39*, 1–97.

Wolfe, E. W., & McVay, A. (2010). Rater effects as a function of rater training context. Upper Saddle River, NJ: Pearson. Retrieved from http://www.pearsonassessments.com

Zhao, L., Deek. F., & McHugh, J. (2010). Exploratory inspection—A user-based learning method for improving open source software usability. *Journal of Software Maintenance and Evolution: Research and Practice, 22,* 653-75.