Journal of Writing Assessment Vol. 3, No. 1, pp. 5-30

Portfolio Assessment
Quantification and Community

NORBERT ELLIOT
VLADIMIR BRILLER
KAMAL JOSHI
New Jersey Institute of Technology (NJIT)

This article presents an outcomes assessment model designed to provide programmatic information to shareholders at a comprehensive technological university. Employing a model emphasizing a veridical relationship between quantification and community, we designed a portfolio assessment process that models a unified validity concept. The assessment model, implemented collegially across the undergraduate humanities curriculum, was found to offer integrated evidence about the ability of our students to think critically, to draft and revise their work, and to document sources used in their assignments.

In 1996, Davida Charney called for composition researchers "to undertake the hard task of inter-connecting our work, by building up provisional confidence in our methods and our knowledge base by challenging and impressing each other—and anyone else who cares to look" (p. 591). In seeming to answer that call, Bob Broad (2000, 2003) presented empirical work to document qualitatively the complexities of programmatic writing assessment through his analysis of more than 700 pages of observational notes, transcripts of group discussions and inter-

Norbert Elliot is Professor of English at New Jersey Institute of Technology (NJIT). His most recent book is *On a Scale: A Social History of Writing Assessment in America* (New York: Peter Lang, 2005). **Vladimir Briller** is Director of Strategic Planning and Institutional Research at Pratt Institute. He also directs educational research projects in Central and Eastern Europe and Central Asia for the U.S. State Department, World Bank, Asian Development Bank and private foundations. **Kamal Joshi** is the Database Manager for Institutional Research and Planning and for Human Resources at NJIT. He holds graduate degrees in both statistics and computer science; he is currently pursuing a Ph.D. in computer science.

Direct all correspondence to: Norbert Elliot, New Jersey Institute of Technology, University Heights, Newark, NJ 07102-1982, elliot@njit.edu.

views, and program documents. His findings document the mismatches that occur in a writing assessment community between the few samples of student work selected for norming purposes and the many student samples to be read (the crisis of textual representation); his research also explores the tensions that arise when readers cannot align their criterion-based evaluative responses to the shifting reading demands found within individual student samples (the crisis of evaluative subjectivity). Although Broad's work is both interesting and important, his data collection and analysis are qualitative. No such studies exist in terms of communities who are committed to quantitative writing assessment. Little is known about the ontological and epistemological orientation of group members, and even less is known about the process of information reification that occurs as research is planned, gathered, interpreted, analyzed, and reported. The philosophical orientation and quantitative results we report here are offered to initiate discussion regarding the veridical relationship between quantification and community.

Background: The NJIT Assessment Community

The "each-other-and-anyone-else" described in the present study exists within a Department of Humanities at New Jersey Institute of Technology (NJIT), a comprehensive public technological university located in Newark. The department consists of 6 professors, 6 associate professors, 3 assistant professors, and 17 lecturers who work together in the service of the university's undergraduate General University Requirements (GUR) shown in Table 1. Although most writing assessment focuses on work done in the first-year curriculum taught by specialists in English, other instructors—holding advanced degrees in anthropology, history, philosophy, and policy studies—host classes across the entire NJIT humanities undergraduate curriculum. Totaling 18 credits, the GUR is understood as writing intensive. That is, instructors of each course within each cohort encourage writing as a cognitively oriented process, one that has grown out of many purposes, that exists in an interrelated fashion with reading and speaking and technology, and that improves with drafting and revision (Writing Study Group, 2004). Anticipating White's (2005) recommendation that a set of curricular aims established by the faculty is essential to Phase 2 portfolio scoring—innovative systems that allows relatively rapid and responsible evaluation, yielding reliable results at a reasonable cost in time and effort, and providing direct information to the faculty on the outcomes of their programs—we established goals for the GUR in 1996. In order to foster an environment in which these goals can be implemented effectively, departmental administrators have maintained class sizes in accordance with standards established by the National Council of Teachers of English (NCTE College Section, 1987).

The department is dedicated to delegating evaluative leadership to instructors. Our assessment system rests on the belief that only a student's course instructor can bear witness to an individual student's command of writing ability, and so the work we describe does not exist within a system in which submitted student work serves as a basis for a high-stakes, pass/fail evaluation. In 1996, the department adopted a portfolio system in order to capture a range of student writing ability

Table 1. General University Requirements in the Humanities

Cohort	**First Year**		**Sophomore Year**	**Junior Year**		**Senior Year**
Course Title	Writing, Thinking, Speaking[a] (basic writing)	Writing, Thinking, Speaking (traditional composition)	Cultural History[b]	Electives in Literature, History, Philosophy, or Science, Technology, and Society[c]	Open Electives, including Theater and Technical Writing[d]	Capstone Seminars[e]
Credits	6	3	6	3	3	3
Fall 2004 Enrollment	182	414	477	184	206	289
Average Class size	16	23	28	30	29	18

[a]Goal: To provide instruction in written and oral communication in the context of the first-year curricula; to emphasize critical thinking as essential in producing effective expository writing, with readings and writing assignments drawn from the humanities, engineering, and the social and natural sciences.

[b]Goal: To compare and contrast world cultures; to utilize case studies focusing on differing forms of material culture, belief systems, aesthetic norms, and artistic productions to develop an understanding of ancient and modern worldviews.

[c]Goal: To allow students to examine broad issues in the humanities through a wide variety of survey courses.

[d]Goal: To allow students to examine technical and professional issues associated with the humanities.

[e]Goal: To engage each student as a unique individual capable of humanistic appreciation of cultures and their diverse complexities, to engage that student in the course content through seminar techniques; to improve the communications skills of each by means of writing-as-process techniques that reinforce engagement with the course content; to improve the communication skills of each student by means of oral presentation techniques—student-led discussion topics, informal presentations, and formal presentations—that reinforce engagement with the course content.

Note: Equation 1 provides a basic formula for calculating a sample size at a level of confidence within ±E units of the population mean in each of the cohorts:

$$n = \left[Z_{\alpha/2} * \sigma / E \right]^2 \qquad (1)$$

where

$Z_{\alpha/2}$ = 1.65, the Z -value associated with a 90% confidence interval

α = the Type-1 error rate

σ =1.31, the population estimated standard deviation; the estimates in this example, shown in Equation 2, are based the calculations from the overall portfolio score standard deviation from the previous (Fall 2003) semester senior seminars

E = the margin of error, in this case .25, the standard error as calculated from the overall portfolio score from the previous (Fall 2003) semester senior seminars

Hence, inserting the values associated with the Fall 2003 senior seminars in order to calculate the Fall 2004 sampling plan, we find that

$$n = \left[1.65 * 1.31 / .25 \right]^2 \qquad (2)$$

n = 75

within the GUR and to report that ability to our shareholders. From 1996 to 2003, portfolios were used by individual teachers within their classes to afford students a way to document their progress and reflect on their gains. From within each portfolio, instructor and student selected a best paper, often the semester's most complex essay, for submission to a reading conducted by instructors. A portion of these essays, identified through sampling plans such as that discussed later, was then read holistically, the instructors using a traditional 6-point scale. Readings were independent, and a combined score (12, the highest to 2, the lowest) was recorded.

Inter-reader reliability was always acceptable, with few best papers requiring adjudication because readers had failed to award matching or adjacent scores. The Fall 2003 first-year writing adjudicated scores, for example, yielded reliability coefficients of .847 (established by Cronbach's α) and .778 ($p < .001$, established by Pearson's r); the Fall 2003 senior seminar adjudicated scores demonstrated reliability coefficients of .858 (Cronbach's α) and .752 ($p < .001$). In Spring 2003, however, the course instructors voiced concern that insufficient information about student writing performance was being provided by scores gained from best paper readings. For Fall 2003 first-year best papers, the mean score was 7.29 ($n = 55$, range, 4–11, $SD = 1.8$); for Fall 2004 senior best papers, the means score was 8.06 ($n = 45$, range, 3–12, $SD = 2.01$). Instructors had agreed to determine a combined score of 7—a score of 4 from the upper end of the scale combined with a score of 3 from the lower end of the scale—as the minimum desirable score, and so the students were meeting minimum competency standards. But what, the instructors asked, did these single numbers reveal about the ability of the students to think critically, to extend their writing beyond summary into persuasive analysis? Did the students draft and successfully revise their work before submitting final copy? Were students able to cite sources in a standard format, the documentation itself revealing that students were using the ideas of others to inform their submitted assignments? Was there evidence in the portfolios that the students were making oral presentations? Could students work collaboratively? Could an overall portfolio score be established that was holistic in nature and not a mere sum of answers to other questions?

Quantification and Community

Those of us at NJIT who have been associated with providing answers to such challenges—individual instructors interested in assessment, each of the program directors for the four cohorts of courses, and members of the Office of Institutional Research—have tacitly developed an orientation that guides our research. Following Glenn Tinder (1995), we strive for a sense of veracity that extends beyond general definitions of truthfulness to encompass an applied vision of truthfulness about self. Such an attitude toward veracity, as Tinder argued, leads to a greater sense of potential, to the possibility of a created self. Taken thus, veracity allows a sense of hopefulness in which attention is focused on the emerging present and a promising future. Veracity is not an end in itself but, rather, a tie that binds the concepts of quantification and of community.

We associate quantification with empiricism. With Karl Popper (1935/2002), we follow the aphorism of all empiricists—that falsiability is a criterion of demarcation—and thus work to "expose to falsification, in every conceivable way, the system to be tested" (p. 20). Following Kenneth Burke (1945/1969), we take the information gained through such efforts to be knowledge of conditions and relations (p. 194). With Janice M. Lauer and J. William Asher (1988), we celebrate the connections—of investigation, of description, and of persuasion—that exist between empirical and rhetorical inquiry (pp. 4-6). Because quantitative measures are veridical when used in an informed fashion, they are understood to have the advantages noted by Charney: They are open to public and private scrutiny, they are reliable and, thus, can be learned and shared; and they are formal and, thus, can overcome geographic and temporal distance, disparities of experience and background, and absence of a shared natural language (p. 577). We understand quantitative measures not as final, mechanical statements of captured past proficiency but as tolerant, organic representations in the present and future quest for construct representation.

Yet if veracity and quantification may be straightforwardly defined as they apply to a group dedicated to quantitative research, the concept of community is not so easily applied. Indeed, the term is historically elusive when used in association with writing assessment. Troubled by the reductionistic nature of proficiency examinations at their university, Peter Elbow and Pat Belanoff concluded in 1986 that "the only way to bring a bit of trustworthiness to grading is to get teachers negotiating together in a community to make collaborative judgments" (p. 338). Nineteen years later, reflecting on the outcomes statement for first-year composition students formulated by the Council of Writing Program Administrators, Kathleen Blake Yancey (2005) cited Robert D. Putnam's study (2000) of the collapse and revival of American community and wondered what kinds of cross-institutional communities could be formed to enhance student development in a collaborative fashion (p. 219-220). Yet the high-stakes assessment readings described by Elbow and Belanoff are not ideal places for communal reflection—although, as Broad (2000) has demonstrated, they are ideal occasions for the study of crisis. Similarly, Yancey's call for collaborative proposals, however hopeful, is made to a profession whose members would nod in agreement with Reinhold Niebuhr (1932) that moral excellence is reached not by groups but by individuals and who would mumble, for good measure, that such is perhaps the case for research excellence as well. Nevertheless, we have found, with George S. Wood and Juan C. Judikis (2002), that our quantitative community has developed around an assumption of mutual responsibility, an acknowledgment of interconnectedness, and a commitment to integrity that has developed around a common purpose. We turn now those common purposes, what we have termed the five validation goals of our community

Validation Goal 1: The Environment

The heuristic value of metaphor in the field of evaluation has been demonstrated by Nick L. Smith (1981). Although extended study of the perceptual value gained from fields as diverse as law and watercolor painting has been examined by Smith and his colleagues, the concept of sustainability offered by

the environmental field (World Commission on Environment and Development, 1987) lends an additional metaphor to a view of validity as a unitary concept (American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME], 1999, p. 11). Meeting the needs of the present assessment community without compromising the ability of members of that community to meet their own instructional needs has always been a goal of our assessment. In 2003, we realized that the only way to answer questions regarding student ability to think critically, draft meaningfully, document appropriately, present orally, and work collaboratively was to abandon the best paper reading and have the instructors themselves read student portfolios. We knew that the portfolios themselves should capture work done in class and avoid restrictions on content (e.g., Elbow & Belanoff, 1986; Lynne, 2004; Moss, 1992; Murphy & Underwood, 2000; Yancey, 1992). In that our program assessment was designed to preserve a naturalistic instructional environment, issues of topic design that impact tests of writing ability (Breland, Kubota, Nickerson, Trapani, & Walker, 2004) were, although not obviated, viewed rather as part of potential sacrifices in reliability that might reduce construct irrelevance or construct underrepresentation. Such an attitude toward assessment allowed us to follow the belief of Grant Wiggins (1994) that authentic assessment should improve authentic performance and yield insight into such performance, not merely audit it. The achievement of assessment validity, therefore, had to attend to sustainability demands of instruction.

Validation Goal 2: The Content of the Measures

Generally, our instructors accepted the definition of writing expressed by Roberta Camp (1992) as "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" (p. 135). Specifically, that definition was extended so that writing would also be considered (a) an act of critical thinking that extends beyond summary into persuasion, (b) an act that improves with drafting, (c) an act that is best when informed by the voices of others, (d) an act that is interrelated with demands of informal and formal oral presentation, and (e) an act that is often undertaken in collaboration with fellow writers. These five variables—critical thinking, drafting, documentation, oral presentation, and collaboration—thus constituted the content domain of the model, allowing us to address the second goal of our community.

Although the overall portfolio score was to be holistically scored, the five variables were designed to be analytically scored. Thus, dissimilar to the decision of Sara Warshauer Freedman and William S. Robinson (1982) to use only holistic scoring in their program at San Francisco State because of their interest in making global decisions, we sought a diagnosis of student ability for each of the five variables. We have found that analytic reading takes into account the specific features of writing in relation to a general framework, an advantage over the general impressions offered by holistic scoring and the task-dependent orientation of primary-

trait scoring (Purves, Gorman, & Takala, 1988). As a framework, analytic scoring allows the freedom of variation found in sections of courses offered across a 4-year curriculum while allowing readers to focus on the independent variables at hand. Alan C. Purves and his colleagues dealt with an international research project, yet the rationale for his adoption of analytic scoring methods may be taken as metaphoric: The form of a text has both language-specific and language-transcendent qualities. In the case at hand, the variables exist across courses, but the way that an argument may be framed in an introductory section of first-year composition may be vastly different from the implicit argument presented in a senior documentary study project. Hence, analytic scoring allowed the preservation of the five independent variables while allowing for individual variation. Although Brian Huot (1990) has suggested that holistic scoring is usually recommended over analytic scoring, especially for large testing populations, our sampling plans allowed us to use smaller samples, and we desired the precision that analytic scores would yield regarding student performance. We also were hopeful that our NJIT scoring method would yield the association found between analytic and holistic ratings (Veal & Hudson, 1983). Because our model did not seek to extract factors of writing ability, it might be said to resemble what Breland (1983a, 1983b) termed a focused holistic scale—a measure that asked readers to focus on distinct aspects of writing but, in doing so, does not exclude any specific characteristics.

The scoring sheet shown in Fig. 1 was designed to allow readers, first, to form a holistic impression of the portfolio they were reading in each of the four cohorts of courses and then to evaluate analytically each of the five independent variables. Thus, the internal consistency of the end-of-semester portfolio assessment episode was designed to be expressed by the independent (predictor) variables as they are associated with the dependent (outcome) variable of the overall portfolio score shown in Fig. 2. Because no one was willing to state that one variable was more important to a student's writing ability than any other, a weighted variable model was rejected.

No single set of isolated statements found on a rubric can ever capture the complexities of the construct of writing (Broad, 2003). Thus at least three 1-hour meetings are held each semester with instructors working in courses identified in Table 1 to discuss the subtleties of the independent variables shown in Fig. 2 as they are encountered in the classroom. Readings are held at the end of each fall and spring semester for students in the first-year and senior-year courses, as well as for students taking electives in technical writing (Johnson, 2006). Readings for students enrolled in other junior-level electives and for sophomores in cultural history courses are undertaken every other year. In each reading, a range of portfolios is selected to represent scores that are appropriate to each of the independent variables and to the overall portfolio score, and at least 1 hour is devoted to discussion (a term we prefer to norming) of these sample portfolios before the reading begins.

Validation Goal 3: Concurrent Relationships

Although the content is designed to be established by the internal consistency of the five variables and their association with the overall portfolio score, we felt the portfolio scores should also be examined for their con-

Portfolio Assessment

Reader's Name: _____
Student's Name: _____
Course: _____

Holistic Score

Provide an overall, *holistic* impression of the portfolio you are reading.

The materials in the portfolio demonstrate superior work in the class.	The materials in the portfolio demonstrate very good work in the class.	The materials in the portfolio demonstrate average work in the class.	The materials in the portfolio demonstrate below average work in the class.	The materials in the portfolio demonstrate work that is at a level of near failure in the class.	The materials in the portfolio demonstrate work that is at a level of failure in the class.

Overall Portfolio Score

Analytic Scores

Provide an *analytic* reading in which you focus on each of the five traits identified below.

1. The contents of the portfolio demonstrate that the student has thought critically about the course subject matter as described in the syllabus.

Very strongly agree	Strongly agree	Agree	Disagree	Strongly Disagree	Very Strongly Disagree

2. The contents of the portfolio demonstrate that the student has drafted and successfully revised papers before final copies were submitted.

Very strongly agree	Strongly agree	Agree	Disagree	Strongly Disagree	Very Strongly Disagree

3. The contents of the portfolio demonstrate that the student can site and document sources by using a standard format (e.g., Modern Language Association format, the Chicago Manual of Style, or the American Psychological Association format).

Very strongly agree	Strongly agree	Agree	Disagree	Strongly Disagree	Very Strongly Disagree

4. The contents of the portfolio demonstrate that the student has made oral presentations in class.

Very strongly agree	Strongly agree	Agree	Disagree	Strongly Disagree	Very Strongly Disagree

5. The contents of the portfolio demonstrate that during the course the student has had experience working in teams

Very strongly agree	Strongly agree	Agree	Disagree	Strongly Disagree	Very Strongly Disagree

Figure 1. NJIT Portfolio Assessment Scoring Sheet

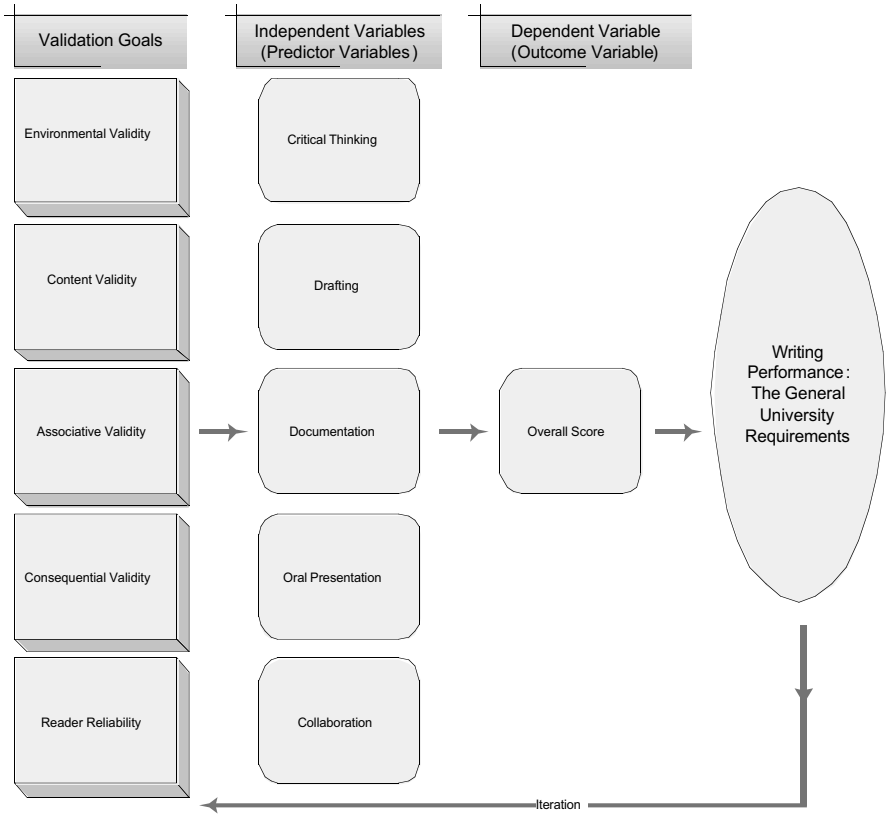

Figure 2. NJIT Portfolio Assessment Model

current relationships with other measures. This third validation goal is defined in the present model as the relationship of the variables to admissions tests (the SAT Reasoning Tests in mathematical and verbal ability used before the 2005 College Board revisions) and to placement tests (in reading, sentence sense, and essay performance, all tests based on forms of the New Jersey Basic Skills Placement Test). As a measure of its concurrent relationships, the portfolio scores were also examined for their association with a holistically scored best paper from each course (retained in 2003 and withdrawn in 2004), as well as the grade in the course that the portfolio was designed to capture. As a measure of their predictive power, the portfolio scores were examined for their association with the student's grade point average (GPA) calculated the semester following the assessment episode.

Validation Goal 4: Consequences

To judge their consequences, portfolio scores were used to analyze student performance. As a measure of consequential validity, student scores were examined across administrations to see if significant differences existed and to interpret the reasons for such differences.

Validation Goal 5: Reader Reliability

The final validation goal, reader reliability, was designed to establish measures of inter-reader agreement and inter-reader reliability. Although promising new scoring systems are beginning to emerge (Ostheimer & White, 2005), questions of inter-reader reliability remain largely unanswered in portfolio assessment (Broad, 1994, 2000; Callahan, 1995, 1997; Koretz, Stecher, Klein, McCaffrey, 1994). In the NJIT program, both inter-reader agreement and inter-reader reliability are held as essential qualities that are preconditions to analysis for the validation goals of environmental, content, associative, and consequential validity. As such, any trait described in Fig. 1 that received scores that were not adjacent (i.e., 6 and 4), was read by a third reader for adjudication purposes, a procedure often termed the parity model (Johnson, Penny, Fisher, & Kuhs, 2003). All observations were made independently, and instructors did not read the portfolios of their own students. Within this environment, inter-reader agreement and reliability were understood as valid evidence affording the assessment of stated curricular goals. As Moss (1994) suggested, evidence of reliability is put on the table for discussion as part of a comprehensive system designed to reflect a range of educational goals (p. 10; see also Williamson, 1994). Reliability is thus part of a network on information that ranges from instructor chats in hallway conversations to the scores we report in statistical tables.

These five validation goals constitute the core values of our quantitative assessment community. Certainly, they are not to be taken as isolated concepts but, rather, as propositions offered to support a unitary concept of validity. Messick (1994) argued, and we agree, that special validity criteria need not be established for performance assessments and that evaluation of consequences must be part of all general validity standards. Following Messick, we designed the validation goals of our community to ensure adherence to validity standards regarding content, substantive, structural, external, generalizability, and consequential aspects of the construct of writing ability. As the following results demonstrate, we believe the quantitative results support (i.e., lend veracity to) the validation goals of our assessment community, thus establishing a veridical relationship between quantification and community. Within our model, the absence or presence of veracity should be understood as the degree to which the accumulated, integrated evidence supports the interpretation of portfolio scores as they reflect a unitary concept: the effectiveness of a cohort of humanities courses within the NJIT GUR as these courses provide instruction in writing.

Quantification: Toward Validation

There is great interest in the writing skills of entering and exiting undergraduate students. Thus the results that follow focus on students taking first-year courses and senior seminars during assessment episodes in Fall 2003 and 2004. The discussion that follows is intended to demonstrate the benefits of quantitative study that have been realized by our community.

Our performance assessment seeks to gain information about the program's effectiveness and not about individual students. As such, it is not necessary to evaluate the portfolio of every student, although university policy requires that each student maintain a portfolio for each course identified in Table 1. Because analytic portfolio evaluation is more time consuming than traditional holistic evaluation, fewer, randomly selected portfolios are desirable. Equation 1 shown in the note to Table 1 provides the formula we currently use in sampling plan design. Thus, although 289 students remained in the senior seminars after the withdrawal date, only 75 student portfolios were needed in Fall 2004 to yield a 90% confidence level. The students whose portfolios were not selected for the collaborative reading reviewed the portfolios individually with their instructors during office hours.

Once the total target number is established each semester, the program developer obtains a class list for each section of the class from the university's Student Information System (SIS). Students within each section are assigned numbers (in this case, from 1 to 289), and a table of random numbers is used to select students who must then submit their portfolios to their instructors for collaborative evaluation. To validate the representative accuracy of the sampling plan, comparisons are made between the students within the sampling plan (n) and the total student population under consideration (N). In the case of students targeted for collaborative portfolio review in Fall 2003 ($n = 55$) and 2004 ($n = 60$), the sampling plan captured a representative group of total first-year writing students ($N = 698$ in 2003, $N = 596$ in 2004). In the case of senior students targeted for collaborative portfolio review in Fall 2004 ($n = 80$), the sampling plan captured a representative group of total senior-year writing students ($N = 289$). The sampling plan expressed in Equation 1 reflects—within demands of cost and time that prohibit the entire population of entering first-year and graduating students to be studied—the total population of NJIT students within each course cohort. Additionally, the sampling plan has proven sensitive to shifts within student samples, rewarding researchers with smaller samples when standard deviation and standard error are narrow and demanding larger samples when these measures increase. Because our aim is program assessment—not individual student testing—the findings that follow should be understood as descriptions of the effectiveness of the curriculum depicted in Table 1.

Validation Result 1: The Environment

Our program evolved for 7 years before portfolios were read by the instructors and was never understood to be an exit exam. Thus it faced none of the tensions described by Broad (2000). From Fall 1996 to Fall 2003, instructors kept portfolios as vehicles of student reflection, meeting with their stu-

dents at the end of the semester to review their work and to help them select the single sample—their best paper—that would best represent completed work in the course. Because the best paper score had no influence whatsoever on the course grade—or on the instructor's employment future—best paper readings were taken as celebratory.

When the portfolios themselves were eventually brought forward for assessment, required was evidence that students had thought critically about the course subject matter, drafted essays meaningfully, documented sources appropriately, presented information about the course subject matter orally, and worked collaboratively. Our curriculum spreads across 4 years. Thus it would have been intrusive in 2003 to specify that students would have to submit specific assignments, as they were in Broad's study of first-year writers at City University; our assessment orientation is more similar, in fact, to that adopted at Washington State University (WSU 2006), where students are asked to submit work that is organized, focused, well developed, and mechanically correct—the more substantial, the better. (Although, again, there is no consequence, as there is as WSU, for a student whose portfolio receives an unsatisfactory score.) Over 7 years, instructors realized that the best papers were almost universally those that were persuasive, that were substantially researched work extending beyond summary, and that employed a number of sources in service to the demonstrated argument. Hence, submitted portfolios articulated the creation and contour of such work—which was that which the instructors wanted to know more about in the first place.

When we analyzed scores from Fall 2003 first-year reading, we found that the overall portfolio score of the first-year students in the 2003 sample correlated with critical thinking (.632), drafting (.458), and documentation (.465), each correlation reaching the significance level of .01. The documentation variable was correlated with both the oral presentation variable (.356, $p < .01$) and the collaboration variable (.267, $p < .01$). Correlations were also identified between the oral presentation and the documentation variable (.519, $p < .01$). After discussing these data with the instructors, there was general dissatisfaction, similar to that documented by Broad (2000, pp. 232-238), with the representation in the portfolios of the oral presentation and the collaboration variables. The correlations appeared promising between these two variables, one that was remarkable because the correlation had captured the fact that many of the oral presentations were collaborative in nature. Yet, our instructors maintained that a set of PowerPoint slides did not authentically reflect the ability of a student to present information orally, and a list of names on an assignment did not reveal anything about team work. After a second semester of readings with these two variables in Spring 2004, the instructors and program developers decided to withdraw these traits from the portfolio assessment; in that many of the first-year instructors in English teach throughout the 4-year curriculum, hallway chats resulted in a decision to withdraw the oral presentation and collaboration variables from all future assessments.

Table 2 shows Fall 2004 scores. The correlation between the overall portfolio score and each of the independent variable scores increased from the previous year's scores. The correlation between critical thinking and the overall score (.868,

$p < .01$) exceeded the .64 correlation documented by L. Ramon Veal and Sally Ann Hudson (1983, p. 291), as the association between analytic and holistic ratings, and each correlation met the .01 significance level documented by these researchers. A similar pattern of association between each variable and the overall portfolio score was also established for Fall 2004 senior seminar portfolio scores, as Table 3 shows.

Validation Result 2: The Content of the Measures

What is the nature of the relationship expressed in the association between the five predictor variables and the outcome variable of the overall portfolio score? That is, is there empirical validation that this independent–dependent variable relationship is robust enough to capture the writing performance that occurred within the first-year writing and senior seminar classes?

A regression analysis of the first-year writing portfolios relating the independent variables (critical thinking, drafting, documentation, oral presentation, and collaboration) to the dependent variable (overall portfolio score) revealed no coefficient of determination ($r^2 = .057$, $F[5, 49] = .59$, $p = .708$) for Fall 2003 first-year writing scores. Regression analysis of the Fall 2004 model, however, revealed a coefficient of determination ($r^2 = .775$, $F[3,56] = 64.172$, $p < .001$) of the relationship of the independent variables (critical thinking, drafting, and documentation) to the independent variable expressed by the overall portfolio score. That is, for Fall 2004 first-year writing portfolios, 77% of the variability of the overall portfolio score represents the proportion of the variation in the dependent variable (i.e., the overall portfolio score) that is explained by the independent variables (i.e., critical thinking, drafting, and documentation).

A regression analysis of the senior seminar writing model relating the independent variables (critical thinking, drafting, documentation, and oral presentation scores) to the dependent variable (overall portfolio score) revealed a low coefficient of determination ($r^2 = .246$, $F[4, 40] = 3.26$, $p = .021$) for the Fall 2003 senior seminar scores. Regression analysis of the Fall 2004 model, however, revealed a higher coefficient of determination ($r^2 = .548$, $F[3, 76] = 32.91$, $p < .001$) of the relationship of the independent variables (critical thinking, drafting, and documentation) to the independent variable expressed by the overall portfolio score. That is, for the Fall 2004 senior seminar portfolios, 55% of the variability of the overall portfolio score represented the proportion of the variation in the dependent variable (i.e., the overall portfolio score) that is explained by the independent variables (i.e., critical thinking, drafting, and documentation).

In both the first-year and senior-year studies, the coefficient of determination improved as variables were withdrawn from the model that had been determined by instructors as inadequately represented in the portfolios. As we show in the discussion of the consequences, we believe that the coefficient of determination also improved as instructors became more confident in making judgments about the quality of submitted work.

Table 2. Associative Analysis: First-Year Writing Portfolio Scores, Fall 2004

Association	1	2	3	4	5	6	7	8	9	10	11
					Fall 2004 First-Year Writing Portfolios (n = 60)						
1. Critical thinking	—	.596**	.571**	.868**	.257	.29*	.071	.278*	.038	.503**	.356**
2. Drafting	.596**	—	.416**	.618**	.043	.247	-.086	.098	-.177	.128	.04
3. Documentation	.571**	.416**	—	.519**	.056	.099	.035	.069	.195	.276*	.042
4. Overall portfolio score	.868**	.618**	.519**	—	.202	.298*	.028	.266*	.016	.472**	.347**
5. SAT: Math	.257	.043	.056	.202	—	.255	.181	.206	-.149	.226	.41**
6. SAT: Verbal	.29*	.247	.009	.298*	.255	—	.420**	.497**	.081	.024	.045
7. Placement: Reading	.071	-.086	.035	.028	.181	.420**	—	.510**	.218	.163	.083
8. Placement: Sentence	.278*	.098	.069	.266*	.206	.497**	.510**	—	.132	.221	.039
9. Placement: Essay	.038	-.177	.195	.016	-.149	.081	.218	.132	—	.128	.029
10. Course grade	.503**	.128	.276*	.472**	.226	.024	.163	.221	.029	—	.643**
11. GPA: Next semester	.356**	.040	.042	.347**	.410**	.045	.083	.039	.029	.643**	—

*p < .05
**p < .01

Table 3. Associative Analysis: Senior Seminar Portfolio Scores, Fall 2004

ASSOCIATION	1	2	3	4	5	6	7	8	9	10	11
					Fall 2004 Senior Seminar Portfolios (n = 80)						
1. Critical Thinking	—	.667**	.636**	.699**	.209	.027	.103	.034	.256	.363**	.206
2. Drafting	.667**	—	.632**	.673**	.25	-.054	.08	.135	.054	.097	.241
3. Documentation	.636**	.632**	—	.526**	.02	-.195	.026	-.187	.054	.161	.304**
4. Overall Portfolio Score	.699**	.673**	.526**	—	.077	.094	.055	.14	.265	.43**	.346**
5. SAT: Math	.209	.25	.02	.077	—	.461**	.075	.148	.214	.083	.201
6. SAT: Verbal	.027	-.054	-.195	.094	.461**	—	-.04	.180	.5*	.354*	.117
7. Placement: Reading	.103	.08	.026	.055	.075	-.04	—	.643**	.091	.094	.125
8. Placement: Sentence	.034	.135	-.187	.14	.148	.180	.643**	—	-.122	.173	.097
9. Placement: Essay	.256	.054	.054	.265	.214	.5**	.091	-.122	—	.378**	.293
10. Course Grade	.363**	.197	.161	.43**	.083	.354**	.094	.173	.378*	—	.463**
11. GPA: Next Semester	.206	.214	.304**	.346**	.201	.117	.125	.097	.293	.463**	—

Note: In the fall of 2003, 29 first-time, full-time senior students in the sampling plan took the SAT Math and Verbal admissions tests, and 32 students took the three NJIT placement tests. In the fall of 2004, 44 first-time, full-time senior students in the sampling plan took the SAT Math and Verbal admissions tests, and 40 students took the three NJIT placement tests. Admission and placement test correlations are based on these numbers.

*p <.05

**p <.01

Validation Result 3: Concurrent Relationships

Beyond the internal relationships of the model, we wanted to know if relationships existed with other measures of student ability. As such, we examined relationships between the model and functional, criterion-based performance levels of the students: our admissions test, the SAT Reasoning Tests in mathematical and verbal ability; and placement tests in reading, sentence sense, and essay performance, designed originally with New Jersey higher education faculty and the Educational Testing Service (ETS). In our analysis of the Fall 2003 first-year portfolio scores, no relationship was identified with either section of the admissions test and the portfolio model. For first-year writers, a lack of relationship was also noted between our model and the placement tests, with only a single correlation identified between the collaboration score and the placement test in reading (.277, $p < .05$). As Table 2 shows, in 2004 there were only weak correlations between the SAT verbal section and the critical thinking scores (.29, $p < .05$) and between the admissions test and the overall portfolio scores (.289, $p < .05$). Low correlations were identified between the sentence section of the placement test and the critical thinking scores (.278, $p < .5$) and between that placement test and the overall portfolio score (.266, $p < .05$). Table 3, describing the senior seminars, also shows no relationship between the model and admissions tests or between the model and placement tests. Our analysis reminds us that admissions tests, placement tests, and performance assessments are intended for different purposes. Although each test may be internally related—note the correlations between the verbal section of the SAT and placement tests in Tables 2 and 3—performance assessments such as our portfolio project seek full construct representation and, thus, may demonstrate little or no relationship with timed tests.

What empirical validation can be provided that demonstrates a relationship with the portfolio model and a holistically scored assessment of the students' best papers and the grade in the course, both measures of concurrent validity? In its earliest form, (2003) the model had no correlation with the best paper when we analyzed the first-year portfolios. However, relationships were identified between the course grade and the critical thinking variable (.335, $p < .05$), the drafting variable (.309, $p < .05$), the documentation variable (.394, $p < .05$), and the overall portfolio score (.394, $p < .01$) In Fall 2004, as shown in Table 2, relationships of increasing statistical significance were identified between the course grade and the critical thinking variable (.503, $p < .01$) and between the course grade and overall portfolio score (.472, $p < .01$). In our senior seminars, a relationship was found in 2003 between the best paper and the overall portfolio score (.378, $p < .01$), although no relationship was found between any variable of the portfolio model and the course grade. As Table 3 demonstrates, however, in 2004, statistically significant relationships were established between the course grade and the critical thinking variable (.363, $p < .01$) and between the course grade and the overall portfolio score (.43, $p < .01$). The existence of these correlations demonstrates that student writing ability, captured in the course grade, may be also captured within the variables of the portfolio model. The moderate range of these correlations also serves as a cautionary reminder that full construct representation exists only in the classroom.

What empirical validation can be provided that demonstrates a relationship with the portfolio model and the students' GPA in the semester following course completion, a measure of predictive power? In 2003, only the documentation variable was associated with the next semester GPA of first-year writers (.339, $p < .01$). In Fall 2004, as Table 2 shows, statistically significant relationships were found between the critical thinking variable and the next semester GPA (.356, $p < .01$), and between the overall portfolio score and the next semester GPA (.347, $p < .01$). In the case of the senior seminars, as Table 3 shows, correlations were established between the documentation variable and the next semester GPA (.304, $p < .01$) and between the overall portfolio score and the next semester GPA (.346, $p < .01$). Although the correlations here are low to moderate, the level of statistical significance is acceptable. However tentative the associations here, predictive patterns exist between the writing ability of students, as defined and captured in portfolios maintained in humanities courses, and the overall academic ability of students in our comprehensive technological university.

Validation Result 4: Consequences

A warranted conceptual shift in educational evaluation toward responsive constructivist evaluation is correctly associated with Egon G. Guba and Yvonna S. Lincoln's (1989) influential *Fourth Generation Evaluation*. Their emphasis on an assessment's credibility, dependability, and confirmability is integrated with their belief that shareholders are often "groups at risk" who stand to "lose their stakes should the evaluation result—in their view—in negative findings" (p. 51). Because our quantitative research is designed to tell us about the performance of our programs and the ways we may improve them in the service of our students, we turn now to a description of the way we have answered two key questions: How did students perform on the assessment? What action was taken by our community once the scores were obtained? These are questions asked by NJIT university administrations who are accountable to accreditation by the Middle States Association of Colleges and Schools (MSACS), the Accreditation Board for Engineering and Technology (ABET), the Association to Advance Collegiate Schools of Business (AACSB) and the National Architectural Accrediting Board (NAAB). Each agency requires an outcomes assessment plan and its implementation. These questions are asked by department administrators and by instructors who want to improve their classes based on the results portfolio review. Because portfolio assessment is not a high-stakes venture for them, they are questions our students never ask—but the way the questions are answered impacts their daily lives as they register for the required and elective humanities courses shown in Table 1.

Within the department, instructors have agreed that scores of 12 and 11 indicate superior work, 10 and 9 indicate very good work, and 8 and 7 indicate average work. Significantly, instructors have also agreed to determine a combined score of 7—a score of 3 from the lower end of the scale combined with a score of 4 from the upper end of the scale—as the minimum desirable score on the analytic assessment of any trait and on the holistic assessment of the portfolios. Any score below 7 sug-

gests below average work and is cause for concern. As Table 4 indicates, none of the scores on critical thinking, drafting, documentation, or the overall portfolio score fall below this cutoff score for first-year students in either 2003 or 2004. Instructors were pleased to learn that first-year students demonstrated competency in each of areas investigated, yet the highest levels of scores in ranges 9 or above are demonstrably absent. Additionally, there was a significant difference in the scores from the first to the second year. Results on an independent sample *t* test, shown in Table 4, demonstrate that there was a significant difference in the scores on critical thinking ($t = 1.94$, $p < .05$), drafting ($t = 3.3$, $p < .01$), and documentation ($t = 2.42$, $p < .05$) variables, although not in the overall score. Although the lower scores recorded in Fall 2004 could reflect comparatively weaker student performance (as measured by the independent variables of critical thinking, drafting, and documentation), there is no indication that the student population shifted. Also, the same instructors taught across scoring periods. A more plausible interpretation is that the readers were becoming more familiar and more confident with the model—as may be inferred by the wider range of scores in the Fall 2004 assessment. This interpretation is supported by the fact that an independent sample *t* test showed that recent Fall 2005 scores were not significantly different from the Fall 2004 scores, whereas the Fall 2005 readers continued to employ a full range of scores: critical thinking (mean = 8.11, range = 5, 11, $SD = 1.38$), drafting (mean = 7, range = 3, 10, $SD = 1.73$), documentation (7.3, range = 2, 10, $SD = 1.69$), overall portfolio score (mean = 7.77, range = 4, 10, $SD = 1.92$). It thus appears that instructors are more willing to use the full range of scores on student portfolios and that the scores have achieved stability.

Indeed, a similar case was evident in the senior seminars. An independent sample *t* test, shown in Table 4, demonstrated that there was a significant difference in the scores on critical thinking ($t = 5.3$, $p < .01$), drafting ($t = 4.97$, $p < .01$), documentation ($t = 3.93$, $p < .01$), and overall score ($t = 2.69$, $p < .08$). The lower scores recorded in Fall 2004 appear to indicate, as in the case with the first-year sample, that the readers are becoming more familiar and more confident with the model. This interpretation is supported by the fact that an independent sample t test showed that Fall 2005 scores for the senior seminars were not significantly different from Fall 2004 scores, whereas—as was the case with the Fall 2004 first-year scores—the readers continued to employ a full range of scores: critical thinking (mean = 7.89, range = 4, 12, SD = 2.20), drafting (mean = 6.59, range = 2, 12, SD = 2.64), documentation (6.27, range = 2, 12, SD = 2.68), overall portfolio score (mean = 7.63, range = 4, 11, SD = 2.03). As was the case with the first-year portfolios, the scores of senior seminar portfolios have achieved stability.

We acknowledge that the low first-year scores in 2003 on the oral presentation trait (6.6) and collaboration trait (6.32) and the low senior seminar scores that year on the oral presentation trait (4.06) might be interpreted as a failure in the classroom and that an unruly faculty conspired to drop those variables from the evaluation. Only the instructors' doubts written on the scoring sheets could assuage such a criticism, and we have no quantitative evidence at present to suggest that these stated curricular goals are, in fact, begin taught. Advancing a sustainability argument, the instructors tell us that they are unwilling to undertake the assess-

Table 4. Score Comparisons: First-Year Writing, Fall 2003, Fall 2004; Senior Seminars, Fall 2003, Fall 2004

First-Year Writing, Fall 2003, Fall 2004

INDICATORS	Range		Mean		Standard deviation		t	P
	Fall 2003 (*n* = 55)/Fall 2004 (*n* = 60)							
1. Critical Thinking	6,11	4,12	8.36	8	1.95	1.68	1.94	.055*
2. Drafting	6,12	3,12	8.4	7.25	1.42	2.91	3.3	.001**
3. Documentation	6,10	2,12	7.94	7.26	.989	1.84	2.42	.017*
4. Oral Presentation[a]	5,9	—	6.6	—	—	—	—	—
5. Collaboration[a]	5,9	—	6.32	—	—	—	—	—
6. Overall Portfolio Score	5,12	3,12	7.29	7.93	1.8	2.04	-1.78	.077

Senior Seminars, Fall 2003, Fall 2004

INDICATORS	Range		Mean		Standard deviation		t	P
	Fall 2003 (*n* = 45)/Fall 2004 (*n* = 80)							
1. Critical Thinking	7,10	3,11	9.22	7.82	1.1	1.55	5.3	.000**
2. Drafting	5,11	2,11	8.99	7.08	1.69	2.13	4.97	.000**
3. Documentation	2,11	2,11	7.88	6.37	1.68	2.32	3.93	.000**
4. Oral Presentation[a]	2,11	—	4.06	—	—	—	—	—
5. Overall Portfolio Score	7,12	2,12	8.88	8.1	1.31	1.7	2.69	.008**

[a]Trait withdrawn from Fall 2004 portfolio assessment.
*p<.05
**p<.01

ment demands that would be necessary to assess quantitatively these variables because such time demands would compromise their own instructional time and the existing assessment program. Such reservations, as we will see here, were not expressed by instructors in reckoning with the low documentation trait scores in the senior seminars.

Validation Results 5: Reader Reliability

Classified by Stemler (2004) as a consensus estimate, inter-reader agreement was solid for the Fall 2003 and Fall 2004 scoring sessions of first-year writing. The lowest percent of agreement was established for the first-year portfolio assessment at 73.3% (for Fall 2004, when nearly 75% of the portfolios needed no adjudication on any trait or on the overall score); the highest was established at 100% (in Fall 2003, in an evaluation of the oral presentation trait). An analysis of inter-reader agreement for Fall 2003 and Fall 2004 senior seminars showed that, when the portfolios were drawn from different disciplines, readers

did not achieve the same uniformity. Just over half (53.3%) of the senior seminar portfolios needed adjudication in Fall 2003. During 2003, the senior instructors also had difficulty judging drafts, achieving only 68.8% agreement in evaluating a variable that was easily handled by the first-year composition readers who had achieved 96.3% agreement. By Fall 2004, however, the senior instructors were able to achieve higher levels of agreement. Seventy-five percent of the portfolios needed no adjudication whatsoever, evidence of an emerging community in which perseverance is its own rewarded virtue.

In reporting inter-reader reliability, we used four forms of analysis, each designed to yield information about the reliability of the reading community. Table 5 provides the reliability coefficients. Inter-reader reliability is reported in terms of nonadjudicated and adjudicated rates as measured by Cronbach's α and Pearson's *r*, both classified by Stemler (2004) as a consistency estimates. Although Cronbach's α provides a general index of reliability, Pearson's *r* allows an estimate of the probability value obtained in a test of significance and a control against Type 1 error (Lauer & Asher, 1988). In that a nonspecific direction of the reliability was assumed (e.g., $\text{Reader}_1 > \text{Reader}_2$ or $\text{Reader}_2 > \text{Reader}_1$) a two-tailed *p*-value was used for the later measure.

With a single exception (that of the assessment of the critical thinking trait in Fall 2003), the nonadjudicated scores of the first-year portfolios yielded a level of agreement, as measured by Cronbach's α, exceeding .519. When analyzed an additional time by Pearson's *r*, Fall 2003 statistically significant correlations were lower (from .351, $p < .01$, to .58, $p < .01$). Clearly, the readers had a difficult time reliably assessing the critical thinking trait. In Fall 2004, however, the patterns of reliability increased for nonadjudicated scores—ranging from .678 to .769 (Cronbach's α) and from .529 ($p < .01$) to .665 ($p < .01$) (Pearson's *r*)—evidence that the readers were evolving as a community. The adjudicated scores achieved, as expected, higher correlations in Fall 2003 and Fall 2004. As was the case with the non-adjudicated scores, the patterns of reliability increased in 2004 for the adjudicated scores—ranging from .789 to .878 (Cronbach's α) and from .657 ($p < .01$) to .783 ($p < .01$) (Pearson's r)—evidence, again, that the readers of first-year writing were evolving as a community. Indeed, in that the score range increased in Fall 2004, as shown in Table 4, the readers demonstrated that they had achieved both agreement and accuracy.

An inter-reader analysis of the senior seminars revealed a similar picture. The nonadjudicated scores yielded a level of agreement, as measured by Cronbach's α, ranging from .24 to .653 in Fall 2003, and only two of the variables met the .01 level of significance, as measured by Pearson's *r*: oral presentation (.501, $p < .01$) and the overall portfolio score (.403, $p < .01$). In Fall 2004, however, the patterns of reliability increased for the nonadjudicated scores—ranging from .599 to .749 (Cronbach's α) and from .444 ($p < .01$) to .599 ($p < .01$)—evidence that the readers of senior seminar portfolios are becoming more cohesive in their judgments. The adjudicated scores achieved, as expected, higher correlations in Fall 2003 and Fall 2004. As was the case with the nonadjudicated scores, the patterns of reliability increased in 2004 for the adjudicated scores—ranging from .741 to .87 (Cronbach's α) and from .599 ($p < .01$) to .771 ($p < .01$)—evidence, again, that the readers of the senior seminars are becoming increasingly cohesive. As was the case with the first-

Table 5. Inter-reader Reliability: First-Year Writing, Fall 2003, Fall 2004;

First-Year Writing								
INDICATORS	**Non-Adj. Cronbach[a]**		**Adj. Cronbach[a]**		**Non-Adj. Pearson r**		**Adj. Pearson r**	
Fall 2003 (n = 55)/Fall 2004 (n = 60)								
1. Critical Thinking	.387	.678	.584	.789	.241	.529**	.415**	.657**
2. Drafting	.734	.783	.767	.878	.58**	.643**	.623**	.783**
3. Documentation	.578	.785	.618	.852	.409**	.651**	.457**	.742**
4. Oral Presentation[a]	.607	—	.607	—	.449**	—	.449**	—
5. Collaboration[a]	.595	—	.702	—	.477**	—	.582**	—
6. Overall Portfolio Score	.519	.769	.702	.827	.351**	.665**	.56**	.708**
Senior Seminars								
INDICATORS	**Non-Adj. Cronbach[a]**		**Adj. Cronbach[a]**		**Non-Adj. Pearson r**		**Adj. Pearson r**	
Fall 2003 (*n* = 45)/Fall 2004 (*n* = 80)								
1. Critical Thinking	.379	.599	.461	.741	.239	.444**	.305**	.621**
2. Drafting	.24	.738	.796	.87	.139	.586**	.663**	.771**
3. Documentation	.358	.749	.797	.749	.219	.599**	.679**	.599**
4. Oral Presentation[a]	.653	—	.824	—	.501**	—	.705**	—
5. Overall Portfolio Score	.575	.732	.667	.836	.403**	.578**	.502**	.720**

[a]Trait withdrawn from fall 2004 senior seminar portfolio assessment.
**$p < .01$ (2-tailed)

year portfolios, the score range increased in the senior seminars in fall 2004; in that reliability was maintained, it is clear that the readers were becoming increasingly capable of accuracy across the 6-point scale.

Truth and Consequences

The work we present here is informed by a unified concept of validity. "The bridge or connective tissue that sustains this unified view of validity," Messick (1989) wrote, "is the meaningfulness or trustworthiness in interpretability of the test source, which is the goal of construct validation" (p. 8). As a form of evidence, Messick believed that social consequences have implications for both the science and ethics of assessment (p. 11). Studies by Huot (2002), Broad (2003), and Lynne (2004) are at one in their treatment of the importance of consequences for writing assessment.

The work presented here—the analytic tables and their interpretation—addresses traditional demands of performance assessment. We are discovering that they are trustworthy, evidence of an evolving community pursuing a cohesive view of assessment. Yet what of meaningfulness? How is the assessment impacting the most important shareholders, the students themselves? There is something comforting about each and every one of the 5,366 undergraduate students at our insti-

tution submitting portfolios to their instructors in 2004–2005, each student reflecting critically on exactly what that course content had meant as papers were drafted, researched, and submitted; each student gathering work to submit a portrait of work accomplished; no student fearing the consequences of that effort. There is something equally comforting about their instructors presenting courses of vastly different content—from textual analysis of Shakespeare's plays to field studies of social programs in Newark—while advancing a coherent vision of writing.

The instructors are the agents driving the program, and so it is appropriate to close our study with a recent initiative based on an extended discussion of Table 4. After we analyzed and presented the data from the Fall 2004 assessment on the documentation trait, there was agreement that the score of 6.37 for senior-year seminar indicated that an important attribute was lacking in the submitted portfolios of graduating students. (Hindsight reveals that there decision was correct; the Fall 2005 documentation trait score of 6.27 would also be unacceptably low.) As our librarian colleagues had been advancing information literacy in the first-year writing course, we asked them to develop an assessment that elaborated the documentation variable. In early Spring of 2005, we designed an information literacy model that associated a student's ability—to identify an original source, to perform independent research beyond the syllabus, to use sources appropriately, and to integrate sources (the independent variables)—with an overall information literacy score (the dependent variable) (Scharf, Elliott, Huey, Briller, & Joshi, in press). Inviting instructors interested in learning more about information literacy, the librarians then conducted a reading of their own. A review of 100 portfolios using our combined analytic and holistic methods yielded higher reliability coefficients than any we had recorded. Adjudicated weighted Kappa coefficients, and additional test of reliability, ranged from .758 ($p < .01$) to .813 ($p < .01$). As well, the model revealed a stronger coefficient of determination than any we had witnessed ($r^2 = .909$, $F[4, 95]) = 238.051$, $p < .001$). Clearly, the model was trustworthy. Yet what the model revealed confirmed our worst fears. Scores for each of the variables were lower for each of the information literacy variables than we had witnessed in a decade's worth of assessment experience: original source (range: 2, 12, mean = 6.68, $SD = 3.01$), independent research (range: 2, 12, mean = 6.46, $SD = 3.25$), appropriateness (range: 2, 12, mean = 6.24, $SD = 3.01$), integration (range: 2, 12, mean = 6.05, $SD = 2.86$), overall information literacy portfolio score (range: 2, 12, mean = 6.14, $SD = 2.9$). When we shared our findings with the instructors—expressing reservations about a single reading needing replication—they listened patiently, yet were convinced that the assessment had captured what they had been telling us all along: Our students lacked important information literacy skills. Although they were unwilling to spend additional evaluative time on capturing the ability of our students to present information orally and to work collaboratively, our instructors were enthusiastic about strengthening connections between assessing and teaching information literacy skills. In that we have presently achieved reliable readings and stable scores in the first-year and senior-year courses, we are in an excellent position to evaluate the impact of curricular innovation.

At the present writing, our findings have been submitted to a university-wide information literacy task force that has been formed by the provost and led by our

dean. Our consensus building project on information literacy is similar to that described by Carol Rutz and Jacquelyn Lauer-Glebov (2005) at Carleton College. Within our department, we are holding a second semester of instruction in information literacy for the first-year class and designing an information literacy curriculum for the senior seminars. Across the years, the assessment has continued to yield such consequential benefits for its community of shareholders. Efficient in its design, the process provides meaningful information to instructors and answers accountability demands of university administrators and accrediting agencies. Most significantly, the process authentically supports student learning, the most significant consequence for anyone assessing writing in an academic community. As such, it serves our camp well, affording a certain amount of light at dusk by which we can tell our stories to ourselves—and anyone else who cares to look and listen.

Acknowledgment

We thank the following members of our assessment community for their continued support: Fadi Deek, Dean, College of Science and Liberal Arts; Robert E. Lynch, Chair, Department of Humanities; Robert S. Friedman, Associate Chair; John M. Coakley, Director of First-Year Writing; Burt Kimmelman, Director of Cultural History; Carol S. Johnson, Director of Technical Writing; and Davida Scharf, Research Librarian. The authors would also like to thank Brian Huot and the reviewers for their thoughtful comments.

References

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

Black, L., Daiker, A., Sommers, J., & Stygall, G. (Eds.) (1994). *New directions in portfolio assessment: Reflective practice, critical theory and large-scale scoring*. Portsmouth, NH: Boynton/Cook.

Breland, H. (1983a). *The direct assessment of writing skill: A measurement review*. (College Board Rep. No. 83-6). New York: College Entrance Examination Board.

Breland, H. (1983b). *Linear models of writing assessment* (ETS Research Rpt.). Princeton, NJ: Educational Testing Service.

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

Broad, B. (1994). "Portfolio scoring": A contradiction in terms. In L. Black, A. Daiker, J. Sommers, & G. Stygall (Eds.), *New directions in portfolio assessment: Reflective practice, critical theory, and large-scale scoring* (pp. 263-276). Portsmouth, NH: Boynton/Cook.

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

Broad, B. (2003). What we really value: Beyond rubrics in teaching and assessing writing. Logan: Utah State University Press.

Burke, K. (1969). *A grammar of motives*. Berkeley: University of California Press. (Original work published 1945)

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. 135-157). New York: Modern Language Association.

Callahan, S. (1995). Portfolio expectations: Possibilities and limits. *Assessing Writing 2*(2), 117–151.

Callahan, S. (1997). Tests worth taking?: Using portfolios for accountability in Kentucky. *Research in the Teaching of English, 31*(3), 295–336.

Charney, D. (1996). Empiricism is not a four letter word. *College Composition and Communication, 47*(4), 567–93.

Freedman, S. W., & Robinson, W. S. (1982) Testing proficiency in writing at San Francisco State University. *College Composition and Communication, 33*(4), 393–398.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*(2), 237–236.

Johnson, C. S. (2006). A decade of research: Assessing change in the technical communication classroom using portfolios. *Journal of Technical Writing and Communication, 36*(4), 413-431.

Johnson, R. L., Penny, J., Fisher, S., Kuhs, R. (2003). Score resolution: An investigation of the reliability and validity of resolved scores. *Applied Measurement in Education, 16*(4), 299–322.

Koretz, D., Stecher, B., Klein, S., & McCafferty, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice, 13*(3), 5–16.

Lauer, J. M., & Asher, J. W. (1988). *Composition research: Empirical design*s. New York & Oxford: Oxford University Press.

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

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*(3), 229-258.

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

Murphy, S., & Underwood, T. (2000). *Portfolio practices: Lessons from schools, districts and states.* Norwood, MA: Christopher Gordon.

National Council of Teachers of English College Section. (1987). *Statement on class size and teacher workload: College*. Retrieved February 16, 2006, from http://www.ncte.org/about/over/positions/level/coll/107626.htm

Niebuhr, R. (1932). *Moral man and immoral society: A study in ethics and politics*. New York: Scribner's.

Ostheimer, M. W., & White, E. M. (2005). Portfolio assessment in an American engineering college. *Assessing Writing, 10*(1), 61–73.

Popper, K. (2002). *The logic of scientific discovery*. London & New York: Routledge. (Original work published 1935)

Purves, A.C., Gorman, T. P., & Takala, S. (1988). The development of the scoring scheme and scales. In T. P. Gorman, A. C. Purves, & R. E. Degenhart (Eds.), *The IEA study of written composition I: The international writing tasks and scoring scales* (pp. 41-58). Oxford: Pergamon.

Putnam, R. D. (2002). *Bowling alone: The collapse and revival of American community*. New York: Simon & Schuster.

Rutz C., & Lauer-Glebov, J. (2005). Assessment and innovation: One darn thing leads to another. *Assessing Writing, 10*, 80–99.

Scharf, D., Elliot, N., Huey, H., Briller, V., & Joshi, K. (in press). Direct assessment of information literacy using writing portfolios. *The Journal of Academic Librarianship*.

Smith, N. L. (Eds.). (1981). *Metaphors for evaluation: Sources of new methods*. Beverly Hills & London: Sage.

Stemler, Steven E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4). Retrieved February 16, 2006 from http://PAREonline.net/getvn.asp?v=9&n=4

Tinder, G. (1995) *Tolerance and community*. Columbia & London: University of Missouri Press.

Veal, L. R., & Hudson, S. A. (1983). Direct and indirect measures of large-scale evaluation of writing. *Research in the Teaching of English, 17*(3), 290–296.

Washington State University. (2006). *Home of the WSU junior writing portfolio*. Retrieved February 16, 2006, from http://www.wsu.edu/~jrpf/

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

Williamson, M. (1994). The worship of efficiency: Untangling theoretical and practical considerations in writing assessment. *Assessing Writing 1*(2), 147–73.

Wood, G. S., Jr., & Judikis, J. C. (2002). *Conversations on community theory*. West Lafayette, IN: Purdue University Press.

World Commission on Environment and Development. (1987). *Our common future.* New York: Oxford University Press.

Writing Study Group of the NCTE Executive Committee. (2004, November). *NCTE beliefs about the teaching of writing*. Retrieved February 16, 2006, from http://www.ncte.org/about/over/positions/category/write/118876.htm

Yancey, K. B. (1992). *Portfolios in the writing classroom: An introduction*. Urbana, IL: NCTE.