Neil Pagano, Stephen A. Bernhardt, Dudley Reynolds, Mark Williams, and Matthew Kilian McCurrie

An Inter-Institutional Model for College Writing Assessment

In a FIPSE-funded assessment project, a group of diverse institutions collaborated on developing a common, course-embedded approach to assessing student writing in our first-year writing programs. The results of this assessment project, the processes we developed to assess authentic student writing, and individual institutional perspectives are shared in this article.

A pendulum swinging between demands for reliability and validity—that is how Kathleen Yancey characterizes the history of writing assessment in U.S. higher education. She argues that from the 1950s to the early 1970s, university administrators and faculty demanded reliable processes for placing students into and out of required composition sequences; the result was heavy reliance on indirect, objective "tests." In the 1970s as composition faculty gained expertise in assessment issues and began to argue that students' writing abilities should be directly measured, universities switched to using actual writing samples to achieve the same goals. To promote reliability, however, the samples were written under controlled conditions in response to a uniform prompt that could be holistically scored using a rubric and benchmark papers. The final period identified by Yancey began in the mid 1980s when composition faculty began to argue even more strongly for assessments that could serve programmatic evaluation goals and provide students with greater formative feedback; the instrument of choice in this period was the portfolio. It is this last phase that Brian Huot so eloquently extols when he argues for *(Re)Articulating Writing Assessment for Teaching and Learning*.

Key to understanding this swinging pendulum is knowledge of the audiences and purposes for writing assessment in higher education. In Yancey's history the audiences have two primary interests: effective course placement and meaningful feedback. As we write this article in 2007, however, a *new* audience for writing assessment in U.S. higher education is appearing on the scene (see especially U.S. Department of Education, *Test of Leadership*), an audience whose gold standard for institutional and program effectiveness is "value-addedness." In a business sense, such a model suggests that any business activity should accrue to the bottom line and produce a net gain. Applied to education, all activities, including teaching, should be tied to demonstrable outcomes.[1] Whether such business models effectively serve academic enterprises is a contentious issue.

This new audience has actually been around for some time in pre-university education. In the first issue of the journal *Writing Assessment* (1994), Pamela Moss discussed the then apparent tensions between what Bryk and Hermanson distinguished as "instrumental" and "enlightenment" uses for assessment in the public school system (quotations from Bryk and Hermanson):

> In the "instrumental use" model, the goals are: to develop a comprehensive set of outcome measures; to examine the relationship between these outcomes and indicators of school resources and processes; and, based upon that generalized knowledge, to control schools through allocation of resources, rewards and sanctions, and regulation so as to maximize performance on the outcomes. . . . An "enlightenment model" of information use reflects a view of schools where interaction among individuals is fundamental and reform requires "changing the values and tacit understandings that ground these interactions (p. 453)." (123)

Moss goes on to characterize the rhetoric of the educational reform movements of the 1990s as instrumental, a prediction that seems to have been fulfilled in the No Child Left Behind Act.

What is new is the application of this agenda to higher education. Comparing different accountability measures, Stephen Klein and his colleagues provide this contextualization of the current situation:

> Over the past decade, state legislatures have increased the pressure on colleges and universities to become more accountable for student learning. This pressure stems from several sources, including decreasing graduation rates and increasing demand, cost, time-to-degree, economic return and public concern for higher education. Moreover, the federal government has placed student learning as one of the top priorities for college accrediting agencies. (251–52)

The new interest in accountability measures based on value-added constructs is likely to have a significant impact on how writing is assessed in U.S. colleges and universities, not only because composition programs will be asked to demonstrate their contributions to student learning, but also because many universities are turning to writing assessments to measure a wide range of learning outcomes, such as critical thinking, research skills, and diversity awareness. As this new wave of writing assessment develops, the trick will be to determine better ways to assess individual change and establish effectiveness relative to national norms, or if not in terms of norms, at least with results contextualized beyond the local institution. Most important will be to build on the principles established in previous waves of assessment history that foregrounded concerns for reliable processes, localized validity, and formative feedback.

This article describes a pilot endeavor undertaken by six institutions with such goals in mind. It begins with a description of a collegial process for jointly assessing authentic, classroom-produced samples of student writing. This description is followed by a discussion of how the assessment findings were used both to gauge student development and to create a public argument for the multiplicity of forces that shape an individual's writing and an institution's writing program. Finally, we consider the project's implications for both the institutions involved and writing assessment more generally.

We present some of the data that was gathered during a FIPSE (Fund for the Improvement of Post-Secondary Education) project. These data provide confirmation that practices across institutions can be usefully compared and that programs within each institution are adding value—they are producing desired outcomes in student writing and faculty communication. However, our emphasis in this report is not on the data per se but on the processes of inter-institutional assessment and the rewards it brings to participating schools.

Needs and Options

As detailed by Klein and his colleagues, after more than ten years of increasing pressure for assessments that contextualize student learning for public accountability purposes, institutions still struggle to find useful ways to assess student achievement, especially in writing. Typically, an institution is faced with two alternatives for direct measurement: the standardized, nationally normed instruments offered by national testing services such as ETS and ACT (characteristic of Yancey's first and second waves of writing assessment) or locally developed instruments (i.e., "homegrown") that are more typical of her third wave.

Each approach presents advantages and disadvantages. Standardized measures, such as ACT's Collegiate Assessment of Academic Proficiency (CAAP) or ETS's Major Field Tests, allow for an institution to compare itself to national means or to those of peer institutions. In addition, these instruments are the result of intense statistical oversight, so issues of internal validity and reliability (from a psychometric perspective) are addressed. However, standardized instruments do not elicit the same type of responses that students produce in class-based work. They are administered in standardized, timed, artificial environments. In addition, student motivation might be questioned: what incentives encourage students to perform their best? Finally, there is the issue of matching the goals of the instrument to the goals of the program being assessed. Very often, these goals do not match very well.

As described by Huot, locally developed instruments typically match the specific goals of the programs they are assessing. In addition, because they often elicit authentic student work as the assessment artifacts (via actual exams and assignments), students' work truly reflects their levels of achievement. However, any benefits of national or inter-institutional comparison are lost when one assesses locally. It is useful to know how an institution is performing within some larger, comparative context, to have some sort of benchmark for contrasting outcomes across institutions. In saying this, it is important to recognize that the value of comparison does not have to reside in positivistic, statistical norms. As described by Huot, collaborative dialogue over what constitutes student achievement nurtures validity and lays the basis for an approach to reliability grounded in consensus (see also Stemler). When that collaborative dialogue reflects inter-institutional collaboration, it also meets the standards set down by the Department of Education's recent report, *A Test for Leadership: Charting the Future of Higher Education*:

> The collection of data from public institutions allowing meaningful interstate comparison of student learning should be encouraged and implemented in all states. By using assessments of adult literacy, licensure, graduate and professional school exams, and specially administered tests of general intellectual skills, state policymakers can make valid interstate comparisons of student learning and identify shortcomings as well as best practices. The federal government should provide financial support for this initiative. (23)

If it is not preferable that we teach in isolation, it is also not preferable that we assess in isolation; collaboration and comparison within and across writing programs add value to our work.

This project responded to the need to develop a collaborative, course-embedded approach to assessing student achievement. We hypothesized that if a group of institutions could develop a shared model of assessment for looking at work from their first-year writing courses, we could garner the benefits of comparison otherwise available only from nationally standardized instruments and, at the same time, provide feedback that would be locally useful and meaningful. The core principle of the project was that course-embedded assessment could give institutions useful performance indices that would also allow the benefits of comparison across institutions.

We believe this project is worth reporting for several reasons. Results of efforts to compare writing performance across campuses is scant but inherently intriguing (for one example see Allen). Second, the methods we used are likely to be of interest to schools undertaking writing assessment. Third, the approach to analyzing student performance suggests factors that might lead to gains in writing ability and, equally as interesting, factors that might not contribute to growth. Finally, we acknowledge that the model itself, of institutions working together to define and measure outcomes of writing courses, presents both challenges and benefits to the participating individuals and programs. We offer this as the beginning of a discussion, not a definitive solution.

Project Funding and Participants

The project was funded by FIPSE, with Neil Pagano of Columbia College Chicago the principal investigator. Initially six institutions were involved: Columbia College Chicago, California State University–Long Beach, California State University–Sacramento, Florida Gulf Coast University, Towson University, and the University of Delaware. Participation in the project was solicited via an assessment-related listserv (*Assess,* supported by the University of Kentucky).

Three inter-institutional assessment projects were originally proposed to unfold over three years: one on writing, a second on quantitative reasoning, and a third addressing diversity. In subsequent revisions to the scope of work, it was agreed to spend the first two years on writing and the third on diversity. The broad goal of the project was to create and test means of assessment that would provide useful comparative data across campuses and useful information on general education outcomes. This report covers only the writing portion of the grant; the initiative on diversity is still under way.

Six institutions elected to participate in the writing part of the project. (Berea College, the seventh institution, chose to participate in the quantitative reasoning and diversity phases of the project.) In the first phase, an initial writing sample was collected toward the end of the first term of first-year composition courses at the different institutions, with a rubric for assessing the samples subsequently developed through a process of collaborative dialogue (described below). After the first year, Towson University dropped out because of campus assessment priorities while Columbia College Chicago, California State University–Long Beach, California State University–Sacramento, Florida Gulf Coast University, Towson University, and the University of Delaware continued. In the second year, writing samples were gathered at the beginning and end of the first-term courses and assessed using the previously developed rubric by representatives from each institution.

Overview of the Project Design Process

The project unfolded within a collaborative, consensus-driven design. Each participating institution appointed a faculty representative to the project, someone with expertise in composition studies. Coordinating discussions, under Pagano's leadership, took place via email during the spring semester prior to a project meeting in Chicago in June 2003. At the project meeting, details were hammered out as to what sort of assessment would be attempted, how it would be managed, and what shared understandings would shape the project on the six participating campuses.

Design decisions represented consensus in some instances and judicious compromise in others. The institutions represented were diverse, with differing missions, student populations, and expectations for student writing. The same could be said of the composition programs that were part of each institution. To the extent that we could, we agreed on what to do and how to proceed. The general principle throughout the design process was to respect the autonomy of individual programs, recognizing that the goals of writing as taught within a particular institutional setting ought to take precedence over any controls or constraints that might be imposed by the project. This point will be returned to at other places in this report. The attempt to embed broadly similar writing assessments into diverse institutional and programmatic contexts proved formidable.

What emerged from this initial project meeting were the parameters for a shared prompt, "Response to a text," a six-trait rubric to assess the student writing, and, through norming sessions, an agreement on standards for the rubric (Appendix 1). Individual faculty at each institution shaped the prompt so it would fit their specific composition sections and could be worked into the instructional sequence of their courses.

During the Fall 2003 semester, faculty in five to eight sections of the first-semester required composition course at the six institutions crafted prompts that fit the previously agreed-upon parameters and assigned the prompt to their students as one of the later assignments in the semester. Table 1 reports the total number of faculty, sections, and student papers from this first phase of the project.

Faculty collected digital copies of student responses to the assignments and sent these and a copy of the assignment prompt to the project director. These were coded, stripped of any institutional identifier, and randomly distributed to the participating faculty for reading and scoring. Each assignment was read by a pair of faculty members, and scores on each category and a composite score (the average of the six categories) were derived.

The group met in Chicago in June 2004 to plan for the second phase. Similar to the first phase, instructors on the participating campuses shaped their prompts within the parameters of a text response. However, this time they created two separate prompts and assigned one early in the term (either the

Table 1: Number of Faculty, Sections, and Students from Project Phase I by Institution

	Faculty	Sections	Students
Columbia College Chicago	3	3	52
CSU–Long Beach	3	3	60
CSU–Sacramento	4	4	69
Florida Gulf Coast University	4	5	75
Towson University	5	5	55
University of Delaware	4	4	61
Total	23	24	372

first or second assignment) and the second later in the term (the penultimate or final assignment). The end result was two sets of student work, responding to similar prompts, from two distinct periods in a semester (one early, one late). Digital copies of the paired samples of student papers were submitted to the project director, who coded and redistributed the pairs of papers to pairs of readers. Therefore, Readers A and B read and rated the pair of papers from Student 1; Readers A and C would read the pair of papers from Student 2, and Readers B and C would read the pair of papers from Student 3, and so on.[2] Similar to Phase I, the prompts were inserted into each file. Also similar to Phase I, readers did not know which institution the papers came from; most importantly, they did not know if the papers they were reading were pre (early) or post (late), and they did not know which papers were from the same student.

Table 2 reports the number of teachers from each institution who were involved in this second phase as well as the total number of sections for which they were responsible and the number of students from which the paired writing samples were collected.

Assignment Prompts

Our search for an assessment prompt that would apply to all of our programs led us to stipulate that the task would involve "responding to a text." As a group we felt that reading, engaging with, and responding to a text in the form of a critical or analytical paper is both a common and important academic task. The parameters were left intentionally broad to accommodate the diverse courses and programs participating in the project. Of note, the term "text" was here understood in its broadest possible context; the definition extended to an essay or editorial, a book, a series of articles, a campus or city environment, or even works of art.

Table 2: Number of Faculty, Sections, and Students in Project Phase II by Institution

	Faculty	Sections	Students
Columbia College Chicago	3	5	66
CSU–Long Beach	6	6	86
CSU–Sacramento	4	4	54
Florida Gulf Coast University	4	5	66
University of Delaware	9	9	107
Total	26	29	379

Each program adapted this central task to their particular situation. Thus, a prompt used at the University of Delaware was worded as follows:

> In his essay "The Human Cost of an Illiterate Society," Jonathan Kozol discusses the problem of illiteracy in the United States. Kozol provides some "real life" examples to illustrate the many difficulties that illiterate people endure in their daily lives. Do you think the circumstances surrounding the individuals he quotes are isolated, or are cases like these more widespread than we would like to believe?
>
> The term "illiterate" can be read as a metaphor for a larger problem beyond the literal meaning of the term itself. In light of Henry's argument in his essay "In Defense of Elitism," what other meaning/s might we attach to the term "illiterate"? How might Henry's idea of "elitism" in education affect or perpetuate this problem? Who do you think is more realistic, Henry or Kozol? Which author do you agree with more? Why? If you are unsure, how might you improve upon their ideas?
>
> I ask that you refer to both the Kozol and Henry essays to support your arguments. Of course, cite your sources according to MLA specifications.

This is a fairly straightforward task, one that might be used at either the beginning or end of the course, and one that represents a central skill as most writing programs are conceived, and certainly so at the University of Delaware, where there is only a single-term first-year writing course, one that stresses academic writing from sources.

This task can be compared to one from Columbia College Chicago, where the program tends to use the surrounding city and its resources as "texts" that its students need to be able to read and respond to analytically and critically:

> At school on the streets of the city:
> Project Three Essay
>
> Please respond to the questions in an essay that includes at least 1000–1200 words of new text (defined as anything that did not appear in the project one or project two essays). I encourage you to revise and re-use any text from previous essays and daily assignments that you think is relevant.
>
> In project one you thought about the character of students; in project two you reflected on the plots that structure students' lives. In project three, then, you will consider the settings of higher education, answering the larger question, what difference does it make that your education happens in this particular place?
>
> As always, I have a few aspects of the problem I want you to consider, and some questions to get you started:
>
> **Histories**—what aspects of Columbia College Chicago's history make it different from the generic image of "College" you explored in project one? How has Columbia College Chicago been shaped by its urban setting, and how has it changed the city? Most importantly, how does knowing more about the history of

> Columbia College Chicago change your sense of the meaning and purpose of a Columbia College Chicago education? (You might use your response to Daily Assignment 3:4 to help you answer these questions.)
>
> **Opportunities**—what can you learn from going to school in the city that you couldn't anywhere else? How does the urban setting change what your education is, and what it means? What lessons are taught to you outside of the classroom, just by the visibility of problems like homelessness and crime, and issues of justice, security, and freedom? (You might use your response to Daily Assignment 3:3 to help you answer these questions.)
>
> **Responsibilities**—if Columbia College Chicago students are, in Jane Jacob's words, "proprietors of the streets" of our neighborhood, what special duties do you have as members of a diverse and complex city landscape? What do these unusual responsibilities add to your education? (You might use your response to Daily Assignment 3:2 to help you answer these questions.)
>
> **Trade-offs**—don't make this essay just an advertisement for Columbia College Chicago. What trade-offs do you make when you decide to attend a nontraditional, urban college rather than a more conventional campus? What burdens or complications result? Do the benefits outweigh the drawbacks? (Any of the Daily Assignments from this project might include useful draft material for responding to these questions.)

In some cases, such as the example from University of Delaware above, the prompts lent themselves well to a pre/post switched-topics design, with half the students responding to a certain text or texts as pretest and the other half to a different text or texts, and then reversing the prompts at the end of term. At other institutions, such as Columbia College Chicago, assignments tended to build off each other, precluding the use of assignments from the beginning of the semester at the end. The particular assignment called upon students to "read" the city in which the college experience is set at Columbia College Chicago. Although this may raise questions from a psychometric standpoint about the comparability of work being sampled, we felt that this was an issue that could be compensated for somewhat by our scoring procedures. As always, we were guided by our interest in embedding assessment prompts into courses that are already part of composition programs, with goals and practices that differ from campus to campus.

The Rubric

The initial rubric developed by the group prior to collecting any writing samples identified six categories, each rated using a five-point scale, with 1 and 2 being low scores, 3 representing adequate performance, and 4 and 5 representing strong performance:

- Evidence of controlling idea
- Engagement with the text
- Organization and development
- Use of source material
- Style
- Command of sentence-level conventions

The general categories identified in this list are best understood by reference to the specific descriptions in the rubric (Appendix 1).

After using this rubric with the samples collected in the first year, we made several changes. We decided it was important and perhaps more reliable to judge whether an essay responded to the task as defined by the assignment than to ask whether the essay was controlled by a strong controlling idea or sense of purpose. These two ways of thinking are similar, but when students are given a well-defined purpose for writing as part of the assignment, we thought it most fair to ask whether the response indeed was responsive to the task as assigned. This allowed us to finesse the issue of students who wrote purposefully but did not write what was asked for in the prompt, something we saw with some frequency during our first assessment round.

We identified a need to separate *organization* and *development* because readers reported encountering texts that were well developed but not organized, or vice versa. We decided to collapse the categories of *style* and *conventions*, since readers were not able to consistently differentiate. We also revised the wording for several of the categories. The five categories for the revised rubric were the following:

- Task responsiveness
- Engagement with the text(s)
- Development
- Organization
- Control of language

Additionally, the scale was broadened from five to six points, primarily so scorers could not default to the midpoint, and a more complete set of descriptors was developed. We confess we have no good explanation for why we went with a five-point scale initially, but during the course of scoring writing samples, we reached the same conclusion others have reached, that a scale with no midpoint is in many ways preferable. Appendix 2 presents the complete final rubric.

The evolution, piloting, and revision of the rubric were a rich process of negotiation, one that tested our individual and institutional philosophies and practices of composition. The categories themselves, stated in simple terms, obscure deeper, underlying differences in conceptualization that emerged during the pilot phase. Could we evaluate a purposeful response, governed by a central idea, or were those terms we used in classes but found difficult to distinguish in samples of writing? Could we reliably distinguish control over language conventions as a separate variable from control over style? What did it mean to engage with the text—what was evidence that a student had grappled with the meaning of a text and was able to discuss some of the complexities of that text and use the text to form some sort of critical or analytical response? Asking such questions encouraged each participant to become more articulate about governing assumptions, expectations, and performance evaluation in our courses and programs.

The Rating Process

In keeping with our desire to follow a process that would meet "accepted" criteria for accountability assessments, we followed standard procedures for norming raters and monitoring inter-rater reliability as outlined by White (*Teaching*) and others (e.g., Myers, Weigle). Individual readers participated in the development of the original rubric at the June 2003 meeting in Chicago and the revision of the rubric in the June 2004 meeting. They also participated in periodic refresher sessions that took place virtually and at CCCC meetings in March 2005. At each of these sessions, the readers discussed student writing samples and matched scores to the rubric descriptors, thereby generating the requisite benchmark papers. There was not complete consistency of readers across the two years. One institution (Towson) decided not to continue in the second year because of local campus priorities for assessment resources. Within the remaining five institutions, four of the five readers were part of the team both years, and a substitution was made at the fifth school, with that new reader integrated through discussion, benchmarking, and practice reading. As a result of this iterative process of consensus building, by the time the early and late writing samples from the second year were rated, the ratings assigned by the two raters for each of the five categories were within one point of each other in 74 percent of the total number of trait ratings (2678/3633).[3] Although some might argue that 74 percent represents borderline acceptability, we would argue that given the inherent difficulty of rating diverse writing samples, it is actually quite high.

An interesting sidebar to this discussion is that although the year-two ratings were conducted without the readers having knowledge of whether a sample was from early or late in the semester, the rater agreement was higher for the writing samples collected later in the semester (70 percent for the early samples, 77 percent for the late).[4] We surmise that this may be because what students produced after experiencing the socializing forces of the first semester composition course was a more homogenized type of writing that was easier for our raters to evaluate; the early writing samples, on the other hand, comprised a more diverse—and thus more difficult to rate—sample. We mention this here as an illustration of how not only the ratings themselves but also the process used to arrive at them can provide a deeper understanding of what occurs in our composition classes.

Collection of Demographic Data

The stated goal in the rhetoric driving current accountability assessment for higher education is the improvement of student learning. Too often, though, accountability assessment stops with the measurement of student learning; a process that simply answers whether students learned. As composition teachers and administrators, we wanted to understand the processes that may—or may not—influence student learning. Because the institutions involved in our project represented a diverse group, we decided to identify parameters along which our programs varied and then examine the relation that different configurations might have to student performance.

The parameters we identified derived from factors related to the institutional context, the instructor (employment status and years of experience), the classroom curriculum (class quota, number of assignments, and typical number of drafts per assignment), and the individual student (SAT verbal score used for admission, gender, intended major, and final grade). Composition teachers know that the list of potential influences is actually much longer. We would have liked to include information about the specific assignments to which students were responding as well as student motivation, for example, but we did not feel that we had reliable ways of defining and collecting data on these factors. We also knew going in that some of these factors that we did define might not generate statistically significant relations to the writing performance scores; nevertheless, we felt that it was important to consider them and report on them as a way of educating the audiences for accountability assessments about the complexities of learning to write.

What Scoring the Essays Told Us

We considered our results in several ways: first, for what they told us about differences between our institutions—the federally desired "interstate comparisons;" second, for what they told us about whether students wrote differently later in the semester than early—the "value-added" perspective; and third, for what they told us about factors that might or might not influence student performance at different times—the locally desired, deeper understanding.

Year One, One Writing Sample Only

Table 3 presents the first-year scores, when each program collected a single sample of writing, late in the term, intended to represent the students' ability to respond critically and analytically to a text.

There were few "surprises" in terms of inter-institutional comparisons. The more selective the institution, the higher the scores. Though there is no simple and objective way to rank institutions on selectivity, college guidebooks include certain characteristics that helped define selectivity. Included in these calculations are acceptance rates and standardized test scores. Table 4 presents the data found in *Peterson's Guide to Four-Year Colleges, 2008* for the participating institutions.

Student papers from the University of Delaware, the most selective institution (based on admittance of applicants and the highest proportion of applicants with ACT composition scores higher than 24) were the most highly rated in all six categories. In fact, when the project had been originally proposed, the project director had hoped to find that student performance might, in fact, be independent of institutional selectivity. This was not the case. Nonetheless, the results suggest with some consistency across institutions that students show reasonable control over the characteristics that underlie the rubric. Most scores fall between 3 and 4, with 3 indicating generally competent performance.

The group learned much from the experience, and as we discussed the results, we came to a consensus about what might prove to be a valuable next step. We decided to replicate this model, yet this time to take two samples of

Table 3: First-Year Mean Scores on Single Sample of Writing, Taken Late in Term, on Six-Category Rubric (standard deviations in parentheses)

	Institution					
	Columbia College Chicago N = 52	California State University–Long Beach N = 60	California State University–Sacramento N = 55	Florida Gulf Coast University N = 75	Towson University N = 42	University of Delaware N = 61
Idea	3.18 (.71)	3.66 (.75)	3.29 (.73)	3.16 (.90)	3.30 (.73)	3.62 (.72)
Engagement	2.92 (.85)	3.45 (.83)	3.08 (.76)	2.93 (.81)	2.80 (.86)	3.64 (.73)
Organization	2.93 (.78)	3.35 (.78)	2.94 (.74)	2.86 (.78)	2.98 (.75)	3.45 (.73)
Source	2.98 (.91)	3.29 (.96)	3.20 (1.25)	2.89 (.96)	2.60 (.87)	3.68 (.87)
Style	3.12 (.72)	3.32 (.70)	3.19 (.77)	2.83 (.79)	3.09 (.80)	3.42 (.77)
Conventions	3.07 (.76)	3.36 (.77)	3.26 (.84)	3.01 (.87)	3.17 (.91)	3.59 (.70)
Composite	3.03 (.63)	3.40 (.72)	3.15 (.71)	2.94 (.77)	3.05 (.73)	3.54 (.66)

Note: Range is 1 to 5, with 1 being weak and 5 being strong. Order of columns is alphabetic. Scoring rubric with descriptors presented in Appendix 1. Composite was not a reader judgment, but simple calculations based on all scores combined.

Table 4: Admissions Selectivity of Project Institutions

Institution	Peterson's Competitive Admissions Rating	Percentage of Applicants Admitted	Percentage of Applicants Whose ACT Composite > 24
Columbia College Chicago	Non-Competitive	95	n/a
CSU–Long Beach	Moderate	52	24
CSU–Sacramento	Moderate	62	12
Florida Gulf Coast University	Moderate	71	22
Towson University	Moderate	69	27
University of Delaware	Moderate	47	73

student writing—one early in the term, one later in the term—and examine them for any possible gains. We had developed the parameters for a shared prompt, a modified rubric with which we were comfortable, a process for reading and rating student work, and the logistics (file coding and sharing, analysis of the results) to make this second phase a promising endeavor.

Year Two, Early and Late Writing Samples (Pre/Post Model)

Tables 5 and 6 present the mean for early (pre-) and late (post-) sample scores by institution, with tests of significance. These means are pair-wise means; in other words, each of the 379 students had both an early sample and a later sample, and mean difference from each sample was compared.

There is remarkable consistency in these data, showing movement in the desired direction at all institutions in every category, with one small exception (a slightly lower late score on "Language" for Columbia College Chicago). What's more, many of the comparisons of early to late scores attain significance at a level of $p < .05$. For each category in the rubric, the average scores for all institutions combined (All) showed significant increases between the early and late essays, probably because significance is easier to achieve given the larger n.

Considering the question that accountability audiences most want answered—do students write "better" at the end of the semester—the findings show statistically significant improvement in scores at each institution. We found this both personally encouraging and a useful result for campus discussions about program value. Aware of comments about the difficulty of documenting student development in writing in the space of a term (e.g., White, *Developing*), we entered this part of our analysis with a healthy skepticism about what we might find. Thus, we were pleased by the consistent evidence of improvement and by the scores tending in the desired direction even when not achieving significance. These findings confirm what Haswell ("Contrasting") has argued: that gains within a semester can be documented if the overall holistic rate is avoided and, instead, if subcriteria are measured. Haswell recommends a method he terms Intrapersonal Paired Comparison. Like Haswell's raters, our raters did not know if they were reading a pre- or post-sample; unlike Haswell's raters, who compared and rated side-by-side pairs of texts from individual students, our raters did not have matched pairs in hand. Nevertheless, our method demonstrated improvement, though perhaps not as much improvement as Haswell's method might have detected.

Although in almost every category at every institution, scores for the later writing samples were higher—and in many cases significantly higher—than the samples from early in the semester, we do not wish to interpret these gains as absolute indicators of what was or was not addressed in the curriculum at individual schools. Writing programs are complex, and we should not load too much weight on a score differential. We do see these scores, however, as a basis for program-internal reflection, allowing each institution to question curricu-

Table 5: Early and Late Mean Scores for Writing Samples with Tests of Significance for Responsiveness, Engagement, and Development

	Responsiveness			Engagement			Development		
Institution (n)	Early	Late	*p*	Early	Late	*p*	Early	Late	*p*
Columbia College (66)	3.76	3.86	.33	3.17	3.47	.00	3.56	3.72	.03
CSU–Long Beach (86)	3.91	4.16	.00	3.75	3.91	.02	3.49	3.73	.00
CSU–Sacramento (54)	3.74	3.78	.75	3.67	3.76	.42	3.60	3.69	.35
Florida Gulf Coast (66)	3.59	3.89	.00	3.37	3.53	.18	3.30	3.50	.06
U. of Delaware (107)	4.13	4.41	.00	3.96	4.24	.00	3.83	4.01	.00
All (379)	3.87	4.07	.00	3.63	3.84	.00	3.58	3.79	.00

Note: Range is 1 to 6, with 1 being weak and 6 being strong. Composite was not a reader judgment, but simple calculations based on all scores combined. Shaded cells indicate statistically significant differences (*p* <.05). Scoring rubric with descriptors is presented in Appendix 2.

Table 6:Early and Late Mean Scores for Writing Samples with Tests of Significance for Organization, Language, and Composite

	Organization			Language			Composite		
Institution (n)	Early	Late	*p*	Early	Late	*p*	Early	Late	*p*
Columbia College (66)	3.51	3.57	.47	3.56	3.51	.57	3.51	3.63	.07
CSU–Long Beach (86)	3.52	3.63	.15	3.52	3.67	.07	3.64	3.82	.00
CSU–Sacramento (54)	3.39	3.46	.47	3.36	3.56	.03	3.55	3.65	.30
Florida Gulf Coast (66)	3.30	3.41	.26	3.41	3.57	.09	3.40	3.58	.04
U. of Delaware (107)	3.81	3.98	.01	3.89	4.14	.00	3.92	4.17	.00
All (379)	3.54	3.66	.00	3.59	3.74	.00	3.64	3.82	.00

Note: Range is 1 to 6, with 1 being weak and 6 being strong. Composite was not a reader judgment, but simple calculations based on all scores combined. Shaded cells indicate statistically significant differences (*p* <.05). Scoring rubric with descriptors is presented in Appendix 2.

lar emphases and instructional practices, as well as to question the biases of the prompts connected with the scores. Discussing the scores among ourselves naturally led us to try to make sense of the differences, to make arguments across campuses, and to think about the prompts and conditions under which assignments were given. We do not believe these scores are fixed or absolute measures; we would expect to see changes in these scores as programs further aligned writing prompts and teaching strategies to the rubric. We also recognize that the variety of conditions and tasks would likely influence the distribution of scores in unpredictable ways, and so we do not place great confidence in the test/retest reliability of these data, especially with regard to comparisons across institutions.

With respect to the inter-institutional comparison, as shown in the final columns of the table, the average of the five-category scores ranked the institutions identically on both the early and late essays, with the institution that has the most selective admission policy (University of Delaware) generating the highest scores, the two California State Universities coming next, and Columbia College Chicago and Florida Gulf Coast University following close behind. As a team we did not find these relative rankings surprising but rather felt that they underscored the importance of considering the institutional context and the background of the students when judging a program—a point we pick up on later.[5]

What Influences Essay Scores?

As noted above, we were interested in exploring the relation between the student performance scores and a group of factors that we had assembled to represent the variety of potential influences on writing, including the institution, the instructor, the class curriculum, and the individual student's background and general expertise. To examine these factors statistically, we used regression analysis, which can evaluate the relative significance of a number of factors simultaneously as predictors of a dependent variable, in this case the student ratings. We used this procedure to study influences on the early and late scores separately. (In this analysis, we looked at only the scores for the second phase of the project.)

For the early scores, we examined factors that we hypothesized might relate to the abilities with which students enter a program, namely the institution where they had been admitted, their SAT verbal scores, their gender, and their intended major. Only the institution proved to have a significant relation. For the late scores, we added factors related to a student's course performance (the overall course grade), the classroom curriculum (enrollment cap, number of required assignments and drafts), and the instructor (faculty/adjunct/teaching assistant and years of experience) to this model. For the late scores the institution again proved to be significant, but we also found that performance on the essays was positively related to the course grade and that students in classes taught by full-time faculty as opposed to adjuncts tended to receive slightly higher scores when other variables in the regression model were held constant. (We had too few graduate assistants in the sample to reach any conclusion.)

We would note that gender, SAT verbal scores, and intended majors were not significant predictors of performance on the essays at either the beginning or end of these initial composition courses. We also did not find a relation between any of our class-related factors (i.e., enrollment cap, number of drafts, number of assignments) and the end-of-semester performance. That we did not find a relation does not mean one does not exist, just that it did not appear in our data. As for what we did find, it is reassuring to know that the end-of-semester performance bears a relation to the final course grade, and for those who believe that the working conditions imposed upon adjunct faculty frequently impact their ability to offer students the same degree of attention and planning afforded by full-time faculty, we offer our findings on faculty status.

Perhaps the most interesting result from this part of our analysis is the dominating influence of the students' institutional contexts. We believe that the proponents of accountability assessment who advocate for "interstate" or "one-size-fits-all" comparisons likely envision the type of comparisons afforded by standardized exams, where institutional results are reported in terms of percentiles. Such percentiles would locate institutions—and students—relative to an anonymous norming population. If students at an institution were labeled as being in the twenty-fifth percentile, the institution would be labeled as serving low-performing students. In our analysis, we found differences between institutions, but we knew who was being compared, and we could make sense of the findings. When it came time to measure progress, we compared our students to themselves, not to national norms or percentile ranks.

This issue of normed populations is key to discussion of any type of accountability assessment. Comparing institutions on normed measures or percentiles is likely to measure input (selectivity) as opposed to value-added (gain). A value-added assessment must account for pre- and post-performance, with an eye on the difference, the value "added" in the equation. Value can be added by teaching performance, but as we frequently acknowledge, the teacher can only do so much in a writing classroom. Improvement in writing skill depends heavily on student engagement and motivation: the student must be willing to work at writing, or there is little likelihood of any value being added. The gains model, based on paired pre/post comparison of performance by the same individual, generates assessment data that are sensitive to learning and skill improvement at the individual level, exactly where we want to be assessing the gain. Note that in this model, students in all institutions have potential—they can improve in comparison to where they started, assuming good teaching but also assuming engagement, motivation, and work on the part of the student.

Programmatic Outcomes of the Inter-Institutional Project

Participation in this project was useful at the local level in that it shaped evaluation practices at program or department levels. With funding and impetus from the FIPSE grant, on each campus groups of instructors met and planned the writing prompt, discussed the rubric and descriptors, and planned for administering a course-embedded writing assessment, on both post-test only and pre/post model. This participation spawned various activities associated with the project but enacted in particular ways at the local level. We highlight here the influence of the project on three participating institutions.

University of Delaware

This inter-institutional project integrated nicely with assessment activities that were being pursued independently at University of Delaware (UD). On our own initiative, we opted for a pre/post model in year one, engaging a dozen instructors in the collaborative research effort. We developed and piloted a set of prompts in two sections, prior to the first year of FIPSE data collection, an activity that led us to consider where students could be expected to be at the beginning of our course and where we would like them to be at the end. We used pre/post switched topics, with two parallel prompts (A and B). Half of the classes wrote on A at the beginning of the term and half on B, with topics reversed at the end of term. The later samples were used for the FIPSE project, but the pre- and post-writing samples were also scored locally by the participating instructors and program administrators. Working with the rubric and refining the descriptors established a feedback loop that encouraged instructors to clarify their goals for their assignments and for the course. Sharing the rubric with students encouraged an approach to teaching where expectations for written work were explicit and categorical.

The project encouraged instructors to articulate their standards, discuss the quality of various essays, and compare themselves to other experienced instructors in the program. A small UD internal grant allowed modest stipends and provided food for a day of norming and scoring the full set of pre- and post-test essays. The group was normed using essays that were unmatched for pre and post, selected to represent a range of quality. The results were of keen interest, with some course sections showing substantial gains, some showing little difference, some suggesting a decline in quality. The effort was manageable, collegial, and enjoyable.

At UD, we acknowledged many threats to reliability within our small assessment project. The students came from different populations, with honors, special needs, and regular admissions students all in the mix. The instructors had different goals for their courses, and not everyone followed the same procedures for embedding the task. There were issues of control over administration of the prompt, with some classes performing strictly in class, some taking the prompt home to write, and some providing opportunities for revision. We put those differences out in the open, and we were careful not to read too much into the results. Treating data as an object of interest helped defuse the threat of using data to compare individual teachers and class performance. Still, everyone was excited to engage in proposing rival hypotheses to explain observed differences. There was much discussion of how student performance is constrained by assignment design and task conditions, and there were spirited arguments about what such results could tell us about our programs.

Having inter-institutional data within a pre/post model proved to be important within the campus climate at UD. We were just engaging as a campus with outcomes assessment, tied to Middle States accreditation review. Our college dean had challenged the writing program to produce evidence that the cost of a large writing program actually produced desired outcomes. His argument was this: "If we are going to spend a large part of the College budget on required writing, we need to know the money is well spent." This comment was not offered in a hostile way; rather, it was a friendly but serious request. We were able to use the FIPSE data with other data we had been collecting, such as a redesigned student evaluation questionnaire targeted at specific outcomes from writing courses. We also evaluated samples of writing from our least qualified students (who were placed into small tutorial sections) alongside the writing of our mainstream students, finding that there were substantial differences in quality and that we ought to maintain our practice of funding smaller tutorial sections for less prepared students. Our dean has since moved to a new institution, but we continue to collect data to make the case for the quality and value of our writing program, and we use the data to make resource arguments (hiring, funding, professional development).

Overall, the FIPSE project stimulated and helped promote a program that sought to measure outcomes as part of its business, with a sense that it could define what was important to measure and how it would be done. The efforts have helped increase budget allocations for program development, outcomes assessment, and faculty training. The ability to benchmark against other campuses helped us think about how our programs and students compare to those on other campuses. Such projects connect us to the larger communities of practice that constitute composition as a field. Working across campuses allowed us to form relationships and to gain perspective because we were able to transcend our local forums of engagement and sites of work.

California State University–Long Beach

Recalling Moss's depiction of "enlightenment" assessment models, which involve a change in values and assumptions among participants, we next turn to California State University–Long Beach. The College of Liberal Arts funded a local extension of the FIPSE project in spring of 2006 by inviting faculty to develop two new rubrics that could be used to assess students' writing—a project that underscores how "value-addedness" complicates assessment of student composition.

First-year writing at California State University–Long Beach is taught in four departments—Asian/Asian-American Studies, Black Studies, Chicano and Latino Studies, and English. Eight faculty members from these departments developed two rubrics for writing assessment by using the FIPSE project as a starting point. Focusing on the category "integration of texts," which is roughly synonymous with the project criterion "engagement with texts," participants early on voiced some disagreement about which samples of student writing best met the category. Some participants valued student writers who were able to demonstrate how the text "relates to" and or "connects with" their personal experiences. Other participants valued student writers who could concisely demonstrate comprehension of source material and then apply that understanding regardless of personal relevance. Two participants identified these difficulties in post-project analyses of their work:

> I was less interested in determining how [students] had "read the problem" and more interested in describing how they had "problematized the reading." And whereas I myself am more likely to value a clumsy counterpoint over a slick summary of the issue, I realize that the University requires us to value both.
>
> I had always conceived of the text in a composition classroom as a sort of stimulus, an "itch," if you will, to which the appropriate response would be a good, long, satisfying scratch. With that in mind, I encouraged my students to use the text as a way of discovering what they themselves thought/felt about a particular

> issue, in other words, to "scratch" the "itch" caused by the ideas in the text, and not necessarily to "integrate" (which I assume to mean "join with?") the text into their writing.

The participants also grappled with constructing a rubric for its two main readers: students and faculty. One participant argued that language in rubrics should include terms that students would immediately understand, such as "phat" and "radical." Other participants objected to this idea, arguing instead for more traditional terms.

An additional complication arose from the nexus between assessing particular student writing and the more general assumptions and values that operate in composition programs. For example, participants were asked to reflect on the program objectives when developing and using rubrics; the program objectives state in part that first-year writing courses should "foster thoughtful questioning of commonplace assumptions" among students. Composition courses should moreover encourage students to understand how writing at college "creates, extends, validates, and amplifies knowledge while at the same time it tests, evaluates and challenges all knowledge claims." Students must consequently "be self-reflexive—learn to inquire into the bases of their own ideas and beliefs in order to compare and contrast their views with the perspectives of others."

The above quotations prompted one participant to question program values. "Whose 'commonplace assumptions' is the Program identifying?" she wrote. "What assessment language will we devise to determine the validity or quality of information that might possibly be outside of the mono-cultural perspective of academia as we know it, and how might the new rubric be rendered less mono-culturally 'centered' or biased?"

While the California State University–Long Beach assessment project ended without full reconciliation of these issues, this participant added value to our work by reminding us that students sometimes develop undervalued interpretations of texts: their logic is unconventional, their syntax is convoluted, etc. She also reminded us that faculty can bring bias to each assessment and that we need to question how personal assumptions may affect public evaluations of student performance. The comparative focus among institutions that the FIPSE project created thus extends to the comparisons individuals can make on one campus: local assessments of student writing can add value to larger projects when faculty are willing to question and revise assumptions about student outcomes. We can be enlightened by readings offered by colleagues across the hall and across the country.

Columbia College Chicago

Columbia College also used the FIPSE project as an opportunity to engage with issues of assessment. The first-year writing program at Columbia College serves approximately 2000 students in 90 sections of three courses (Introduction to College Writing, English Composition I, and English Composition II) each semester, including versions of each course (ESL, Enhanced, Service Learning) tailored to meet the varying needs of a diverse student population. Since English Composition I (Comp I) and English Composition II (Comp II) are the only courses that every student takes, these courses are often seen by administrators as valuable sites for assessment. Unfortunately, our institutional outcomes-based assessment has often left those involved feeling either defensive (reacting to mandates and judgments from people outside the writing program) or passive (just a condition of the jobs of teaching writing and administrating writing programs). Given the faculty's resistance to or grim tolerance of our college's attempt to create a culture of assessment, it was not surprising that the prospect of participating in an inter-institutional assessment of writing failed to interest writing teachers or our composition committee initially. As the composition committee began to look at the completed report, however, they recognized the value in being a part of a larger discourse community on assessing writing.

In the process of bridging our own material locations with others in the study, we realized that program renewal at Columbia College would depend on assessments that made more visible our teaching and intellectual work. In our college's outcomes-based assessment plan, the FIPSE results supported our own judgments about students' writing and demonstrated to administrators our commitment to assessment. While the FIPSE rubric represented six features that were inarguably important criteria for evaluating writing, it could not contain the many other criteria by which a rhetorical performance is judged. In order to achieve a clear and concise rubric, some of the descriptive and informative excess represented in the data had to be ignored. In designing our program's assessment we focused on one point in the FIPSE report as a platform for developing a more comprehensive and detailed picture of our students' writing.

The Composition Committee (director of composition, six full-time faculty, and two part-time faculty) used the report's findings to initiate a discussion of what we valued in our students' writing. We focused on the small drop in the score for conventions from the early sample to the later sample. Committee members offered several ways of accounting for this drop in score, but the consensus was that the shift in score could best be attributed to the more complex kinds of writing we wanted students to be doing by the end of the course.[6] While early writing assignments asked students to respond to texts in standard forms using their own experiences and insights, later assignments asked students to respond from multiple perspectives and to consider form and audience more explicitly. Perhaps the familiar territory of the early writing assignments enabled many students to pay closer attention to the details of copyediting, but as the writing assignments progressed, students devoted more time to experimenting with audience and form. Given the tendency among our students to enjoy tasks that ask them to think creatively, considering form and audience energized the drafting and revision of these later essays. We hypothesized that the time and energy students spent creating innovative forms left less time and mental energy/space for copyediting.

Our interpretation of the FIPSE data led us to design an assessment of our Comp II course that investigated course goals dealing with students' ability to control sentence-level features and demonstrate creativity in using the methods, forms, and genres of research writing. In developing the scoring guide, instructors used the terms "craft" and "creativity" to evaluate these features. Comp II has been the focus of recent curricular revision efforts that have resulted in the development of three inquiry-based themes for Comp II: inquiry, ethnography, and visual rhetoric. While each Comp II theme represents a different focus, in all Comp II courses students develop and sustain a single inquiry over the duration of the course. We gathered a random sample of 300 final researched essays and scored them using a rubric with two categories: craft and creativity. By collaborating on the creation of the rubric, the composition committee was able to learn more about what course goals mean from different perspectives. By assessing these essays the program developed a more comprehensive picture of students' abilities to write creatively and accurately. The breakdown of average scores showed a slight drop in craft score, affirming the FIPSE data that illustrated a similar drop in students' control over sentence-level features; but the breakdown according to the approach of the Comp II section was also instructive. Essays written for the inquiry model of the course, the oldest version of course, was lower than either the overall average or the creativity category but exceeded the average in the craft category. We concluded that repetition has made this version of the course a bit less adventuresome than newer, fresher approaches. However, we also noted that in this course, which places a greater emphasis on more traditional forms of library research than the newer models, students' craft scores were slightly higher.

With average overall scores of 3.225 for craft and 3.765 for creativity on a 6-point scale, both craft and creativity were close to the average. Because the ethnography sections accounted for approximately 60 percent of the overall sample, it exerted substantial influence in the general results. The questions the results posed for the ethnography track echoed the FIPSE study and center on the integration of stronger instruction in the elements of writing craft into the course. Ethnography's lowest-overall craft scores (3.137) may reflect the added difficulty involved in guiding students in the use of a broader range of materials than conventional library research typically includes: site descriptions, transcripts of interviews, etc. Nonetheless, the low score in craft on our local assessment and the low score in control of language in the FIPSE study represent an area that will receive greater attention.

The visual rhetoric model is the newest and smallest area of emphasis in the program, so its results were likely the least reliable and predictive in this sample. Given this track's emphasis on synesthesia and hybrid textual forms, it is appropriate that the work being done there reflects the value we place on innovation and creativity. The very strong results in the craft category also suggested we should perhaps pay some deliberate attention to how the issues of convention and correctness are taught in these sections. Some teachers thought the attention to document design required for the incorporation of visual elements into alphabetic text requires students to pay more strict attention to editing and presentation.

Limitations

We have tried to be circumspect in our claims throughout this report, since in terms of assessment models, we realize our work has limitations. Right from the start, it was evident that we had differences in program philosophies and practices that would lessen our controls on reliability. Composition programs have commitments to certain kinds of writing and writing pedagogies, so it was not easy, or even possible, to standardize a writing task and administer it at all participating sites.

There are fairly straightforward issues of task (or face) validity here. Not all programs were willing to introduce an in-class writing task solely for purposes of assessment, as this was viewed as intrusive and out of character with the purposes of the course. Also, the idea of a self-contained prompt for the writing sample simply did not fit the model of composing that informed the course and therefore would not provide a valid measure of performance. In some programs, students are taught to do research in preparation to write and to take their work through drafts with feedback and revision. To make the project work, we had to surrender control of conditions and task in deference to program philosophy and practices. Thus, we sacrificed reliability (controls) for validity (respecting local practices and philosophies).

In the pre/post model, we faced issues arising from different course designs. Some programs were committed to certain forms of engagement to begin a course, and there was resistance to taking an early sample of writing for assessment purposes. These programs had very specific purposes for early assignments, planned as part of a motivated sequence of assignments and activities. It was not seen as a good idea to insert a kind of writing not typically part of the scope and sequence of the course. Doing so would interfere with normal starting or ending writing activities. Composition programs also tend to have well-formed ideas about how to end courses. Courses tend to build toward culminating, complex forms of writing and engagement, so to say that all the late samples of writing were a simple model of "respond to a text" would be a misrepresentation. We therefore broadened the timing (the first or second assignment, the last or penultimate assignment) and relaxed the controls over conditions. Some students wrote in-class only, some were able to draft and revise, some received feedback, and some did not.

Composition programs tend to work with different genres, based on program design. The task "respond to a text" therefore had to be very broadly defined. Students might be working with or responding to multiple texts, and the kinds of text varied widely. The most interesting variations on the assignments arose at Columbia College, where the course tended toward assignments that sent students into the surrounding environs, and a text might be a museum or a canvas, a theater performance or the city streetscape. Seeing text in such a broad way is theoretically appealing to the semioticians in us all, but practically, with regard to comparability across institutions, such task variability presents problems of comparing very different kinds of writing by a common rubric.

The project, therefore, had serious threats to reliability. We do not believe the value of the project was compromised by the limits on reliability of our data, as long as we are careful about how we interpret the results and what we do with them. We agree with Pamela Moss, who argues that in more hermeneutic approaches to assessment, there can indeed be validity (and value) without strict reliability:

> With respect to generalization across readers, a more hermeneutic approach to assessment would warrant interpretations in a critical dialogue among readers that challenged initial interpretations while privileging interpretations from readers most knowledgeable about the context of assessment. Initial disagreement among readers would not invalidate the assessment; rather, it would provide an impetus for dialogue, debate, and enriched understanding informed by multiple perspectives as interpretations are refined and as decisions or actions are justified. And again, if well documented, it would allow users of the assessment information, including students, parents, and others affected by the results, to become part of the dialogue by evaluating (and challenging) the conclusions for themselves. (9)

As we have tried to show in the descriptions of how the project played out under local conditions on participating campuses, the act of inter-institutional assessment was the occasion for extended discussions about writing, rubrics, scores, teaching, and learning. We have highlighted throughout this report the collateral benefits of the project, and we have attempted to be cautious about making generalizations or inferences on the basis of our data. It is clear, however, that the assessment activities triggered meaningful personal and institutional change.

Conclusions and Closing Comments

In this project, our participating institutions and participants gained new perspectives on assessment, demonstrating in the process that we can do inter-institutional assessment without giving up the things that are important to us. We were able to embed the assessment into existing course structures, allowing for programmatic differences in prompts and conditions. Programs and teachers on each campus took ownership of the assessment initiative and found ways to make assessment work at a local level while still allowing useful comparison and discussion across institutions.

Although some in our field urge us to go beyond rubrics to obtain more richly contextualized assessment data (see especially the arguments in Broad), the development and refinement of the rubric for this project proved a valuable exercise, one that sharpened programmatic goals and classroom practices. It is important that the inter-institutional development of the rubric served as a forum for discussion for those inside the project, while at the same time providing useful data for stakeholders outside the project who are anxious to compare programs and institutions. Broad notes that rubrics are frequently developed without such consultation among involved parties (12). Whether one endorses or finds value in developing, using, and reporting results based on rubrics will always be contingent on the purposes of assessment, the resources available, and the level of granularity or particularity sought in the data. For our purposes, given our project goals and resources, the rubric served the project well.

This project had significant collateral benefits, for programs and for the individuals involved. This inter-institutional assessment project provided an answer to the question of how we are doing, as well as a mechanism for sustaining discussion over a period of time. The project helped us figure out what works and what does not, and it helped us refine important questions about our teaching and the performance of our students. Research typically teaches us the intractability of practice, the irreducibility of complex activity. What is important is often not so much specific findings that answer questions, but the learning that derives from engagement with the process. In that sense, we count this project a success.

Appendix 1: First-Year Rubric and Descriptors for Inter-institutional General Education Assessment Project (IGEAP) Writing Assessment Rubric

	Category	
Low Scores 1 or 2	Average Score 3	High Scores 4 or 5
Evidence of controlling purpose (central idea or argument)		
Fails to establish purpose for writing. No clear point, or purpose; no central argument to paper. Paper drifts substantially from initial purpose or controlling idea.	Purpose or controlling idea is established initially, but inconsistently attended to. Paper shows some unity of purpose, though some material may not be well aligned.	Establishes strong sense of purpose, either explicitly or implicitly. Controlling purpose governs development and organization of the text. Attends to purpose as paper unfolds.

Engagement with the text

Does not connect well to the source text. Does not show evidence of having understood the reading(s) that should inform the paper. Repeats or summarizes source text without analyzing or critiquing.	Shows some evidence that materials were read or analyzed and that those texts have shaped the author's writing. Shows basic understanding and ability to engage the substance of the text(s). Goes beyond repetition or summary of source text(s).	The writer clearly read and understood the source text(s) that inform the paper. Summarizes key points or issues in the source text and then critically analyzes or synthesizes those ideas with the author's own ideas. Extends the ideas of the source text in interesting ways.

Organization and development

Moves in unpredictable sequence. Lacks progression from start through middle to end. Moves from idea to idea without substantial development; lacks depth.	Some evidence of organization, with appropriate moves in the introduction and conclusion and some partitioning in the body. Paragraphs tend to have topic sentences with supporting details. Achieves some depth of discussion.	Establishes clear pattern of development, so the paper feels organized and orderly. Develops ideas in some depth. Uses appropriate strategies of narration, exemplification, and exposition through generalization/support patterning.

Use of source material
(Rate only if the paper uses sources beyond the text that is being summarized and analyzed.)

References to source materials are either not present, or sources are not well introduced. It is often not clear where information is coming from or whether sources are properly cited. In-text citations and end-of-text references are not	Source materials are cited, though not always consistently. It is generally clear when information comes from sources. Most in-text citations have appropriately formatted end-of-text references.	Source materials are introduced, contextualized, and made relevant to the purpose of the paper. It is always clear when information, opinions, or facts come from a source as opposed to coming from the author. Source materials are conventionally docu-

formatted according to an appropriate style sheet.		mented according to academic style (APA, MLA, CSE).
	Style	
Lacks control over sentence structure; difficult to follow; does not use appropriate transitions. Little control over patterns of subordination and coordination. Requires the reader to backtrack to make sense. Uses wrong words and awkward phrasing.	Style is competent, though not engaging or inventive. Shows reasonable command over phrasing and word choice. Some useful transitions and patterns of reference provide connection.	Author clearly controls the pace, rhythm, and variety of sentences. Sentence style is smooth and efficient, with good use of subordination and coordination. Words are well chosen and phrasing is apt and precise. Sentences move smoothly from one to the next, with clear moves that open, develop, and close topics.
	Command of sentence-level conventions	
Many errors of punctuation, spelling, capitalization (mechanics). Many grammatical errors (agreement, tense, case, number, pronoun use).	Some typical errors are in evidence, but overall, the writing is correct.	Few if any errors of punctuation, spelling, capitalization (mechanics). Few if any grammatical errors (agreement, tense, case, number, pronoun use)

Appendix 2: Second-Year Rubric and Descriptors for IGEAP Writing Assessment Rubric

	Category	
Low Scores 1 or 2	Average Scores 3 or 4	High Scores 5 or 6
	Task Responsiveness	
Fails to establish purpose for writing. Does not respond to the task.	Establishes purpose or controlling idea initially, but inconsistently attended to.	Establishes strong sense of purpose congruent with the task, either explicitly or implicitly.

Paper drifts substantially from initial purpose or controlling idea.	Shows some unity of purpose and attention to task.	Controlling purpose governs development and organization of the text. Complicates purpose as paper unfolds.
Engagement with Text(s)		
Shows little evidence of having understood the reading(s) that should inform the paper. Repeats or summarizes source text without analyzing or critiquing. References to source materials are either not present and/or sources are not well introduced. In-text citations and end-of-text references are not formatted according to an appropriate style.	Shows some evidence that materials were read or analyzed and that those texts have shaped the author's writing. Goes beyond repetition or summary of source text(s). Source materials are cited, though not always consistently. It is generally clear when information comes from sources. Most in-text citations and end-of-text references are appropriately cited.	The writer clearly read and understood the source text(s) that inform the paper. Summarizes key points or issues in the source text and then critically analyzes or synthesizes those ideas with the author's own assertions. Introduces, extends and complicates the ideas of the source text. Consistently clear where information, opinions, or facts come from a source as opposed to coming from the author.
Development		
No apparent awareness of readers' needs or expectations. Claims have little or no logical support. Moves from idea to idea without substantial elaboration; lacks depth. Shows little or no support for narrative, analytic, and/or expository patterning.	Some awareness of readers' needs and expectations. Claims are logically supported. Ideas have sufficient elaboration; achieves some depth of discussion. Shows sufficient support for narrative, analytic, and/or expository patterning.	Anticipates readers' reactions to key points in paper. Claims are logically supported with relevant, compelling detail. Ideas are substantially elaborated with significant depth. Shows significant support for narrative, analytic, and/or expository patterning.

Organization		
Paragraphs do not cohere. Topics not clearly introduced, explored, or concluded. Essay moves in unpredictable and illogical sequences for the reader. Lacks progression in form from start through middle to end. Patterns of exposition/analysis/argument unfold with little discernable purpose.	Paragraphs generally cohere with topic ideas and supporting details. Topics are introduced, explored, and concluded. Essay moves in predictable and logical sequences. Shows progression in form from start through middle to end. Patterns of exposition, analysis/argument unfold according to purpose of paper.	Paragraphs cohere and make overall essay more complex. Topics are clearly introduced, explored, and concluded in interesting ways. Essay moves in logically surprising and satisfying ways for the reader. Shows compelling progression of form from start through middle to end. Patterns of analysis/argument anticipate readers' expectations and meet purposes of paper.
Control of Language		
Lacks control over sentence structure; difficult to follow; does not use appropriate transitions. Little control over patterns of subordination and coordination. Requires the reader to backtrack to make sense. Uses wrong words and awkward phrasing. Grammatical errors disrupt a reader's progress through the essay.	Style is competent, though not engaging or inventive. Shows reasonable command over phrasing and word choice. Some useful transitions and patterns of reference provide connection. Uses mostly appropriate words and phrases. Occasional grammatical errors do not disrupt a reader's progress through the essay.	Author clearly controls the pace, rhythm, and variety of sentences. Style is smooth and efficient, with good use of subordination and coordination. Consistently deploys useful transitions and patterns of reference. Words are well chosen and phrasing is apt and precise. Virtually no grammatical errors.

Acknowledgments

We express our thanks to our colleagues who participated in the project but who are not co-authors of this report: Linda Rowland (Florida Gulf Coast University), Amy Heckathorn (California State University–Sacramento), Marsha Watson (Grand Canyon University), Lena Ampadu (Towson University), and Faun Bernbach Evans (Columbia College Chicago). We also thank the CCC *reviewers for careful and constructive review and commentary. We are indebted to FIPSE for funding and flexible support. Finally, we thank the students, teachers, and schools that were such willing and engaged participants in this study.*

Notes

[1] Two relevant positions on "value" include Edward White's 1990 assertion that statisticians aimed for ostensibly "value-free" assessments of student writing ("Language" 197). English departments should consequently "define the value" of our teaching by reconsidering how language works in our discipline and our assessments (196). Theodore Hershberg more recently lauds assessment aims put forth by No Child Left Behind, which promotes "value-added" assessment whereby student performance is separated into the work supplied by students and the work provided by teachers and schools (280).

[2] The method described here is similar to Haswell's (*Contrasting Ways*) paired comparison method.

[3] This figure represents all of the work done by the raters. In some instances only one writing sample from a student was scored. Additionally in some instances raters did not provide scores for all five traits. If five traits for two samples for each of the 379 students had been scored, the total number of cases would have been 3,790. The pattern of greater agreement among the raters on the later samples holds on a trait-by-trait basis as well.

[4] The difference in these proportions is significant at $p < .05$.

[5] A reviewer pointed out that it is also interesting to consider the relation between relative gain during the semester and institutional selectivity. There was not a discernable pattern here, although the University of Delaware showed the highest or next to the highest gain from the early to late writing samples for each trait.

[6] Haswell *(Gaining Ground)* offers a similar explanation for drops in demonstrated command of conventions across time.

Works Cited

Allen, Michael S. "Valuing Differences: Portnet's First Year." *Assessing Writing* 2 (1995): 67–89.

Broad, Bob. *What We Really Value: Beyond Rubrics in Teaching and Assessing Writing*. Logan, UT: Utah State UP, 2003.

Bryk, A. S., and K. M. Hermanson. "Educational Indicator Systems: Observations on Their Structure, Interpretation, and Use." *Review of Research in Education* 19 (1993): 451–84.

Haswell, Richard H. "Contrasting Ways to Appraise Improvement in a Writing Course: Paired Comparison and Holistic." Paper delivered at Conference on College Composition and Communication, St. Louis, MO, April 1988. ERIC Document Reproduction Service, ED 294 215. 28 Sept. 2007 <http://comppile.tamucc.edu/pairedcomparison.htm> Uploaded to CompPile 17 May 2005.

———. "Documenting Improvement in College Writing: A Longitudinal Approach." *Written Communication* 17 (2000): 307–52.

———. *Gaining Ground in College Writing: Tales of Development and Interpretation*. Dallas: Southern Methodist UP, 1991.

Hershberg, Theodore. "Value-Added Assessment and Systemic Reform: A Response to the Challenge of Human Capital Development." *Phi Delta Kappan* 87 (2005): 276–83. L11.P53

Huot, Brian. *(Re)Articulating Writing Assessment for Teaching and Learning*. Logan, UT: Utah State UP, 2002.

Klein, Stephen P., et al. "An Approach to Measuring Cognitive Outcomes across Higher Education Institutions." *Research in Higher Education* 46 (2005): 251–76.

Moss, Pamela A. "Can There Be Validity without Reliability?" *Educational Researcher* 23 (1994): 5–12. JSTOR. U. Delaware, Newark. 27 Sept. 2007.

Myers, Miles. *A Procedure for Writing Assessment and Holistic Scoring*. Urbana, IL: ERIC Clearinghouse on Reading and Communication Skills, National Institute of Education, and NCTE, 1980.

Peterson's Guide to Four-Year Colleges, 2008. Lawrenceville, NJ: Peterson, 2007.

Stemler, Steven E. "A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Inter-Rater Reliability." *Practical Assessment, Research & Evaluation* 9 (2004): 1 Oct. 2007, <http://PAREonline.net/getvn.asp?v=9&n=4>.

U.S. Department of Education. *A Test of Leadership: Charting the Future of U.S. Higher Education*. Washington, D.C., 2006. 1 Oct. 2007. <http://www.ed.gov/about/bdscomm/list/hiedfuture/reports/pre-pub-report.pdf>.

Weigle, Sara Cushing. *Assessing Writing*. Cambridge, UK: Cambridge UP, 2002.

White, Edward M. *Developing Successful College Writing Programs*. San Francisco: Jossey-Bass, 1989.

———. "Language and Reality in Writing Assessment." *College Composition and Communication* 41 (1990): 187–200.

———. *Teaching and Assessing Writing: Recent Advances in Understanding, Evaluating, and Improving Student Performance*. 1985. Rev. ed. San Francisco: Jossey-Bass, 1994.

Yancey, Kathleen Blake. "Looking Back as We Look Forward: Historicizing Writing Assessment." *College Composition and Communication* 50 (1999): 483–503.

Neil Pagano

Neil Pagano was the principle investigator for the FIPSE-funded project on which this article is based. He is associate dean for the School of Liberal Arts and Sciences at Columbia College Chicago and was formerly Columbia's director of assessment.

Stephen A. Bernhardt

Stephen A. Bernhardt is chair of the English Department at the University of Delaware. He holds the Andrew B. Kirkpatrick, Jr. Chair in Writing. He is past president of both the Council for Programs in Technical and Scientific Communication (CPTSC) and the Association of Teachers of Technical Writing (ATTW).

Dudley Reynolds

Dudley Reynolds is an associate teaching professor of English at Carnegie Mellon University's Qatar Campus where he serves as the director of research in English language learning. His research interests include the development and assessment of second language writing abilities, learning outcome assessment, and the teaching of English as an additional language in academic contexts.

Mark T. Williams

Mark T. Williams is an associate professor of English at CSULB, where he coordinates the first-year composition program. His research interests include rhetorical theory/history, basic writing, and ESL issues.

Kilian McCurrie

Kilian McCurrie teaches in the first-year writing program at Columbia College Chicago. He also directs the literacy program and coordinates assessment for the English department.