Writing Assessment in the Humanities: Culture and Methodology

Jason M. Barrett

Journal of Assessment and Institutional Effectiveness, Volume 2, Number 2, 2012, pp. 171-195 (Article)

Published by Penn State University Press

Writing Assessment in the Humanities: Culture and Methodology

JASON M. BARRETT

ABSTRACT

This article examines methodological and institutional challenges for empirically measuring student performance on writing. Writing's intrinsic subjectivity and the great variety of writing formats appropriate to diverse contexts raise fundamental questions about the empirical bias of the assessment culture taking root in U.S. higher education. At the same time, the academic training of humanist scholars, who typically have primary responsibility for writing pedagogy in universities, may predispose them to skepticism about assessment culture's broader mission. This article narrates the process by which the Humanities Department at Lawrence Technological University implemented a writing assessment process designed to address these challenges and evaluates the data generated by this process.

The emphasis in assessment culture upon empirical, quantitative measurements of pedagogical effectiveness poses unique challenges for humanist scholars and faculty. Perhaps these challenges stem from a kind of preconscious psychological disposition. We poets, philosophers, historians, and assorted textualists/culturalists are instinctively skeptical of the objectivity-truth claims made with such breezy confidence by scientists and statisticians. But at the root of the matter is a practical incompatibility between the kinds of intellectual skills we think we are developing in our

JOURNAL OF ASSESSMENT AND INSTITUTIONAL EFFECTIVENESS, Vol. 2, No. 2, 2012

students and the kinds of measurement tools that yield numbers with any plausible claim to objectivity.

A student submits an essay on a course reading. Consider three forms of evaluation I might return to them:

1. 80/100
2. B–
3. Well-written and mostly accurate book report, but does not engage the interpretive issues raised in class discussion.

Which of these three forms are "measurements"? The evaluation that looks most like numbers has the smallest pedagogical value. The assignment does not contain a finite number of discrete, "correct answers," for which "80/100" expresses a proportion scalable to other grading criteria. One hundred- or four-point scales merely establish equidistance between the *qualitative* criteria implicit in the letter grade. Eighty out of one hundred means "low B": somewhat better than competency, not quite proficiency. The least-numerical evaluation is by far the most pedagogically substantive. If this student's next paper is more conceptually sophisticated than his or her first one, it is far more likely to be the effect of evaluation 3 than 1 or 2. But evaluation 3, even though it may be represented as "80/100," does not exist along a linear scale with different "quantities" of the same form. A different paper that was more poorly written but more analytically engaged might also be represented as "80/100," but would be the *equivalent* of the first paper in only the most limited and uninteresting way. There is no straight path through Bloom's taxonomy.

Inured to subjectivity by habit or training, humanists are typically not much troubled by these inconsistencies. Understanding is just one of many things in the world resistant to objectification and best examined through language and metaphor. The merciful among us may create point systems to help ease our students' grade anxiety, but those systems generally reflect a weighting of qualitative evaluations. And so we might happily persevere in a kind of penny-ante poker game with our students to entice them into broadening their minds and developing their self-consciousnesses, were it not for the reports of a coming transformation in the institutional culture of higher education that will play a dramatically higher-stakes game with the points we assign to our students' work. The chairperson of the assessment committee at my university, a colleague from the College of Engineering, assures me that "there is nothing that can be taught that cannot be

measured." My humanist predisposition immediately leads me to ponder how one might go about validating such an assertion. But my instincts as a junior faculty member serving on a powerful, university-wide committee allow me to sublimate that skepticism and become an open receptor to the inverse institutional injunction: you had better be able to measure whatever you claim to be teaching. This injunction in itself would not be especially troubling. As a humanist I can see "measure" as a metaphor for a variety of approaches. But when I look at the three forms of measurement in which I most commonly engage, I notice that only one of them works in an Excel formula.

This article relates the efforts of one small, diverse humanities department to engage proactively with the pedagogical and institutional tensions implicit in this anecdote. The tangible product of that engagement is a writing assessment process that may provide a useful model for other departments faced with increasing accountability for their writing curriculums. The implementation of this process required the resolution of a succession of challenges that were broadly of two types. The methodological challenges derived from the goal of creating an empirical, quantifiable standard by which to measure subjective, qualitative student work. Another set of challenges derived from the institutional culture of the department and its position within the university. The specific assessment machine we have invented would not likely work in any other department. It is the unique concoction of the chemistry of our institution. As a kind of guidebook to others who may wish to construct their own assessment machine, this article will detail the nuts and bolts of our process, provide the methodological rationales for its various features, and discuss some of the institutional issues that had to be accommodated in the course of its implementation.

Institutional Context

Lawrence Technological University (LTU) is a private, suburban Detroit school with a traditional focus upon the technical professions and applied sciences. Roughly two-thirds of LTU's 4500+ students pursue undergraduate degrees. Of the 3000+ undergraduates, more than half are in programs within the College of Engineering and more than one quarter are within the College of Architecture and Design. The bulk of the remainder pursues natural science degrees within the College of Arts and Sciences.

The general pattern of undergraduate curriculum at LTU has students devoting the major portion of their first two years to the College of Arts and Sciences receiving a broad liberal arts education in math, natural science, and the humanities, and the major portion of their final two years within their technical programs. The primary institutional role of the Humanities Department is to deliver this "Core Curriculum" in the first two years.[1] An important secondary role for the department is to maintain programs—particularly in economics, psychology, technical communication and business management—that offer courses required by the professional programs. The eighteen full-time faculty members in the department are thus, necessarily, a polyglot assemblage of specialists: historians, philosophers, literary scholars, psychologists, economists, political scientists, lawyers, MFA writers, and so on.

The pedagogical division of labor within the university thus creates fundamental institutional relationships. The smaller liberal arts departments prepare students for the curriculums in the larger, technical programs. And while the engineering and architecture faculty may have some sense that their students are more interesting and imaginative people for having read Plato and Dante, the technical programs are primarily invested in the humanities curriculum to the extent that it develops their students' ability to write. All seven courses in the humanities' Core Curriculum are writing-intensive, and by custom virtually every course offered by the department has a writing component. Whether we are teaching constitutional law or macroeconomics or modernist poetry, the department standard is that we assign papers and devote class time to writing pedagogy.

This sustained focus is necessitated both by the skill profiles of the students who tend to be drawn to LTU and by the dearth of writing pedagogy outside the Humanities Department. LTU students tend to be those who took extra math and science in their high school preparation, and avoided writing and literature courses when they could. Few arrive at LTU writing well and virtually none writing naturally. Students have a strong sense of this bifurcation in the college-bound track between preparation for fields that require reading and writing, and fields that require math and science. To a remarkable extent, they self-identify as "gear heads" and "byte heads," and this self-identification is cultivated by the cultures of their technical programs. In numerous conversations with me, colleagues from the technical programs have enthusiastically endorsed the notion that writing well is crucial to their graduates' career prospects, acknowledged that they themselves much prefer to read well-written prose than poorly written prose, admitted

that a substantial portion of the course work they assign is submitted in prose form, and yet nevertheless insisted that they do not teach writing and would not presume to grade writing. The art of writing appears to them mysterious, obscured in grammatical arcana. It is the provenance of the humanists.

The Methodological Divide

These institutional and cultural orientations set the context for the assessment of student writing on campus: the *politics* of assessment, we historians are apt to say. A university report on the state of advanced-undergraduates' writing skills, issued in the fall of 2008, activated those politics and elicited the engagement by the Humanities Department alluded to above.[2] The report's assessment was dismal: one-third of the papers sampled received an F, and the remainder registered in descending proportions from D to A. The data indicated that fewer than half of LTU students were coming out of the Core Curriculum with a basic competency in writing. The politics of assessment directed accountability for that problem toward the humanities faculty and curriculum. And the politics of the humanities faculty's response was greatly complicated by the realization that the review panel had based its methodology upon the criteria for grading papers that the Humanities Department itself had created.

In two of its earliest efforts at complying with the new standards of assessment, the Humanities Department had created "Guidelines for Writing Papers" and a "Banned-Error List" in order to codify the learning goals and standards of evaluation for its writing curriculum. The Guidelines set out qualitative descriptions of A, B, and C papers focused primarily upon content and analysis. The Guidelines are formatted as a checklist, so that faculty can mark the descriptors that best apply, attach the Guidelines to the paper, and return to the author as an explanation for the grade received. The Banned-Error List explains five elementary grammar errors—fragments, comma splices, run-ons, subject-verb disagreements, noun-pronoun disagreements—and prescribes a grade penalty ("up to a 1/3-grade reduction") for each type of error in a student's paper. The list is taken to be a minimum; faculty members are encouraged to expand it for their courses.

Humanities Department policy directs all faculty to distribute these documents to their students and apply them as the standards for grading papers, but also acknowledges that each faculty member will interpret the

language of the documents in the way that seems most reasonable to that faculty member. The latter caveat was necessary to satisfy issues of academic freedom that emerged in the course of developing and ratifying the documents. Indeed, the language of the documents reflects diplomacy and consensus building. Since a zero penalty falls within "up to," and since faculty are encouraged to expand the list of errors and so increase the cumulative weight attached to grammar, there really is no way to be in violation of the policy. Actual practice in applying the documents, as one might imagine, has varied widely. Some faculty apply them with zeal (especially in first-year composition), others distribute them with their syllabus and may mention them again when they distribute paper topics weeks later.

The assessment panel started from the entirely reasonable premise that the Guidelines and Banned-Error List represented an accepted standard for evaluating writing and sought to apply them to their samples with a mathematical precision and rigor. They read each sample paper and checked the appropriate descriptors in the Guidelines in order to derive a base A, B, or C score. They then scoured each sample paper for grammatical errors, distinguishing them into two penalty categories. "Major errors" were limited to the five specific rules in the Banned-Error List and carried a penalty of a one-third grade reduction (from B to B—) for every three errors. The report classified all other types of grammar and syntax errors as "minor errors" and applied a one-third grade penalty for every five such errors. The same methodology was applied to every sample paper irrespective of its content, length, or the form of the prose required by the nature of the assignment (the samples having been gathered from every college in the university). Multiple readers of each sample paper compared and collated their lists of errors and a final tally for each category of error was entered, alongside the base score, into a data set. Distinguishing the colleges from which the papers were drawn, in juxtaposition with the three data points generated through the scoring, permitted a series of tables in absolute numbers, percentages, and proportions.

If technology permitted, I would poll readers of this article on their immediate response to the validity of this methodology for writing assessment. In the absence of data, I will forward as a hypothesis that we would find a correlation between credulity/incredulity toward the method and readers' academic training. Statisticians are likely to see a reasonable effort to take a subjective body of material, impose upon it a sequence of substantive categories of evaluation (taken from an authoritative source), and weight them proportionately to their pedagogical value. If the final grade scale appears out of alignment, they might say, adjust the internal metrics by lowering the

criteria for base scores or decreasing the penalty per error. Even if we were able to reach consensus on the letter-grade value of specific rates of grammar errors, humanists are likely to see in this anecdote a nightmare scenario in which evaluation of their pedagogy is hijacked by quantifiers. Even strong partisans of back-to-basics writing pedagogy are likely to be shocked to see the range of their investment in their students' writing reduced to a checklist incrementally diminished by middle-school grammar errors.

LTU's humanities faculty were not constrained in responding to the report's methodology by an inability to explain how the Guidelines and Banned-Error List are typically applied in our courses, or why that application is pedagogically substantive and generally just. Our primary constraint was that all of our explanations for how the panel's literalism had distorted the documents' terms tended toward the conclusion that it was impossible to measure writing skills objectively: that the documents were *only* capable of being applied by individual faculty within their courses. To take one example, it is clearly impossible for faculty unfamiliar with the content of a given course to evaluate the content or analysis of papers written in that course. On what basis did the engineers on the assessment panel determine whether a paper's interpretation of *Moby Dick* was insightful or merely competent? And that principle taken to its logical conclusion might preclude anyone but the actual instructor of a given course from evaluating the content and analysis in students' writing. A student who submitted a transcript of a lecture might well earn high marks from faculty familiar with the material but not privy to the content of class time. Yet to declare the nature of humanities curriculum to be such that each faculty member must measure his or her own effectiveness is to fall considerably short of the objectivity standards set by the dictum: "there is nothing that can be taught than cannot be measured."

The Prevailing Assessment Culture

This awkward circumstance of being compelled to explain why our own documented criteria and point systems were not good metrics for assessing student writing, as well as a genuine curiosity to know whether the underlying epistemological issues could be resolved or at least addressed, led LTU's humanities faculty to embark upon a year-long project in developing a writing assessment process. The first stage of that project required an evaluation of the forms of assessment currently employed within the Humanities Department.

Contemporaneously with the creation of the Guidelines and the Banned-Error List, the department had instituted a writing assessment regimen for its Core Curriculum courses. This regimen required the faculty members assigned to each of the Core courses to assess student writing in those courses on a three-year cycle. The six courses that formed the first- and second-year humanities Core were separated by pairs, each pair to be assessed in annual rotation. In practical terms, this policy required the three to five faculty members who regularly teach first-year composition (for example), to organize themselves every third year, collect copies of every paper submitted in the fall term in every composition section taught, and in winter term to read and evaluate those papers. Depending upon course enrollment and rates of compliance in collecting copies of papers, a typical sample size would be 200–400 papers per course. Each cohort of faculty adopted its own scoring rubric and methodology, but the basic process was similar for each: read and score the papers by the rubric, tally the results, construct some graphs, and write a report explaining the graphs.

The department's assessment practices had thus become partitioned into "silos" of bureaucratic responsibilities. These silos were further partitioned laterally to the extent that none of the faculty cohorts had implemented their process uniformly over successive cycles, established benchmarks, or attempted to measure the effects of curricular reforms between assessment cycles. The starting point for a new assessment process arrived with the realization that these silos had to be broken down and a uniform methodology adopted that could be presented to the rest of the university as the department standard. Those discussions began with a focus upon a common assessment rubric for Core Curriculum courses. In the course of these negotiations, as we discussed how each interpreted and would apply various formulations of competency in various schemas, the methodological possibilities of having the entire faculty serve as a reader pool for papers sampled from every Core Curriculum course became apparent. There was a collective wisdom, if not objectivity, in the sum of our individual pedagogical idiosyncrasies. And this wisdom might be capable of being expressed statistically.

Designing a Writing Assessment Machine I: A Common Rubric

The assessment rubrics developed by each faculty cohort employed various point systems, weightings, and categorical and descriptive language. The composition rubric emphasized structural features; the literature

rubric referred to "interpretation" and "voice" while the history rubric referred to "argument" and "evidence." Such differences were not major obstacles to the development of a common rubric. The root compromise that made the common assessment rubric possible was the identification of three equally fundamental (and hence equally weighted) categories of writing skills: argument, evidence, and mechanics. The common assessment rubric (fig. 1) identifies three fields of evaluation within each category and provides qualitative descriptions of A, B, C, D, and F competencies in each field. A thirteen-point scoring key converts those qualitative descriptions, allowing for "high" and "low" degrees of each letter grade, into numerals.

The three categories of writing represent the faculty's collective sense that the intellectual skills we are cultivating in our students find expression in these three elements of their writing. The primary focus of the Core Curriculum courses is textual exegesis. We expect that as students become more skilled at perceiving and deconstructing the analytical or thematic architecture of the canonical texts, their own writing ought to become more self-conscious about its premises and assertions. Hence, the "argument" fields ask readers to evaluate three dimensions of this analytical self-consciousness. The "evidence" fields of evaluation acknowledge that papers are the primary means by which students report their understanding of course content: that reading is as essential a skill in our courses as

HSSC Writing Assessment Rubric

SCORING KEY	12	11	10	9	8	7	6	5	4	3	2	1	0

	CATEGORIES	"A"	"B"	"C"	"D"	"F"
ARGUMENT	A Thesis, main idea, interpretation	Insightful, original; complete thesis statement and 'roadmap' for body of paper in introduction	Coherent, clear and complete thesis statement, but unambitious; restates consensus in class discussion	Thesis statement addresses assigned topic, but overly general, noncommittal, or restates topic as an assertion	Thesis statement vague, not clearly relevant to assignment	No thesis statement
	B Development of argument through body	Body logically unfolds claims in thesis, with increasing conceptual nuance; Skillful use of concession and qualification	Body logically unfolds central claims in thesis, but lacks nuance, concessions or qualifications	Body sustains theme/topic of thesis, but not in an analytically sequential manner	Substantial portions of argument of questionable relevance to thesis	Arguments irrelevant to thesis
	C Counter-arguments anticipated	Takes 'other side' into account and gives strong reasons for author's approach.	Skillful but inconsistent consideration of reasonable counter-arguments	Counter-arguments introduced, but issues unaddressed or unresolved	Counter-arguments undermine thesis	No counter-arguments acknowledged
EVIDENCE	D Command of course material	Demonstrates mastery; material subjected to critical analysis	Demonstrates proficiency; identifies and accurately explains relevant passages	Demonstrates competence; college-level interpretation that does no violence to the text	Coherent, but excessively vague or general	Confused about basic issues
	E Relation between evidence and claims	Each primary claim and many secondary claims supported by direct textual evidence	Each primary claim supported by plausible and relevant examples from texts	Textual evidence mixed with opinion	Relies primarily on opinion	Relies nearly exclusively on opinion
	F Citation	All necessary citations provided, all in proper format (MLA/APA)	Most necessary citations provided, all in proper format	Inconsistent citations, all from valid sources	Insufficient citations	Few or no citations
MECHANICS	G Concision / Style	Cannot be substantially cut; succinct, direct, active prose	Largely free of errors of style; some colloquialisms, but little repetition or redundancy, few passive constructions	Avoids errors of style that impair meaning, but prose is colloquial, repetitive, or passive	Essay could be substantially shortened; mechanical errors impair meaning	Majority of text is superfluous
	H Grammar / syntax	No significant, basic errors	Few and incidental grammar and syntax errors: does not repeat same error	Moderate frequency of errors, or same errors repeated, but meaning unimpaired	Mechanical errors impair meaning	Pervasive mechanical errors

FIG 1. HSSC Writing Assessment Rubric

writing, and papers are the primary way we can give credit to students who are strong readers even if they are comparatively weak writers.

The "mechanics" fields of evaluation represent the most important compromise among the faculty that enabled the common rubric to gain consensus. The base assertion of the rubric is that writing mechanics are a fully co-equal branch of writing skills with analysis and content. Among the humanities faculty, this is a victory for the banned-error partisans. Yet the specific kind of elementary grammar errors included in the Banned-Error List are reduced to one of three fields of evaluation, and hence but one-ninth of the total score for each paper. Stylistic elements and paragraph structure are added as concessions to those faculty members who consider the banned-error rule Draconian or pedagogically narrow-minded.

The qualitative language in the descriptions of A, B, C, D, and F competencies was exceedingly malleable. There are endless combinations of intensifiers and Bloom's taxonomical terms. Broadly, the language tries to depict a mastery and self-consciousness of skill for the A descriptions, a proficiency and thoroughness for the B descriptions, a competency and compliance with minimal requirements for the C descriptions, an active engagement with debilitating faults for the D descriptions, and a lack of engagement or awareness for the F descriptions. All this language is intrinsically subjective. But the fundamental purpose of the descriptions is to mark clear gradations in competency within each field. The pertinent question for the validity of the instrument is not "Would you use these phrases in evaluating a B paper?" Rather, it is "Is the phraseology of the B descriptions clearly distinguishable from those of the A and C descriptions?" To the extent that the rubric succeeds in this purpose, we may at least say that the language of the rubric is one of the least-subjective elements of the assessment process.

Designing a Writing Assessment Machine II: Sampling and Scoring Methodology

Sensitivity to this notion that each course within the Humanities Department's Core Curriculum requires distinct forms of analysis lay at the base of the assessment "silos" that had developed within the department. The composition, literature, and history cohorts were more likely to find consensus in their evaluative terms. The institutional imperative to establish a unified writing assessment process pushed LTU's humanities faculty

toward inverting this methodological problem. The root insight was to abandon the quest for uniformity and instead seek to maximize the diversity of perspectives evaluating the samples, and then establish statistical norms out of the resulting data.

The eighteen full-time faculty members in the department appeared to present a ready-made pool of scorers exponentially larger than any in current practice (this is to say, they were draft-eligible by the service clauses in their contracts.). But the real prospects for unanimous compliance with additional assessment responsibilities would depend upon a substantive transformation in the institutional culture of the department. This change would require the curricular-turf concessions necessary to dismantle administrative silos. More daunting, it would require an active, conscientious participation by the vast majority of a group whose general attitudes toward assessment culture may fairly be described as running from resigned acceptance to principled hostility. A majority or even determined minority might well be able to impose the policy, but it would require something close to unanimity of purpose to make the process actually work.

Initial negotiations centered upon a reasonable workload for each faculty member to assume on an annual basis. Under the former policy, one in every three years the majority of humanities faculty (those who teach Core Curriculum courses) faced an enormous time commitment for assessment. This commitment included reading and scoring 100 or more sample papers, and also committee time in organizing the process, collecting samples, agreeing upon results, and writing reports. The minority of faculty who do not teach Core Curriculum courses are typically responsible for assessment activities within their programs, and these responsibilities would not be diminished by their participation in writing assessment for the Core. Out of these considerations, a proposal was developed that each faculty member be required to score one lot of twenty sample papers every academic year, and that this scoring would constitute their full assessment obligations toward the department's writing curriculum. An administrator ("PA") would be appointed to implement all other aspects of the process, who in turn would be relieved of other service obligations.

The consensus that developed around this proposal was the single most important prerequisite to the success of the experiment. Without unanimous faculty "buy in" to the process, the theoretical possibilities of a different approach to writing assessment would have remained practical impossibilities. The twenty-paper quota was a substantial net reduction for faculty accustomed to the triennial assessment marathon; yet they

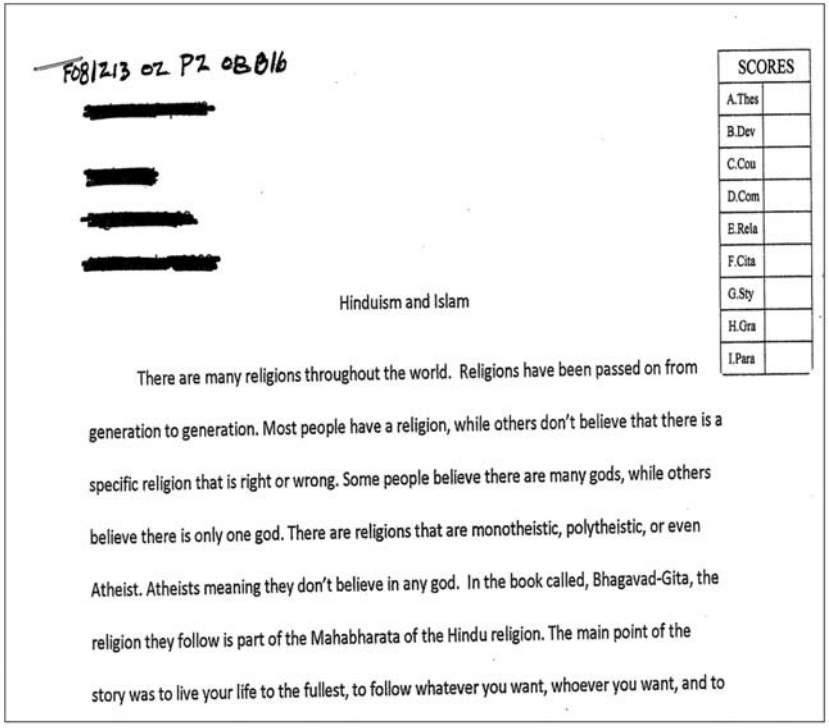
also made the greatest concessions in terms of control over evaluation of their curriculum. The willingness of those faculty outside the Core to take on an additional assessment chore can only be accounted to their remarkably generous spirits and the strong esprit-de-corps within the department. A large majority of the faculty had become invested in the success of the project: some from a genuine curiosity about the empirical outcomes, many more from a sense that the project represented a constructive and politically astute response to the fiasco of the university assessment committee's report on advanced undergraduates' writing skills.

Agreement on a quota permitted the rest of the process to unfold, for we knew that the assessment machine we were constructing could produce a finite number of scored samples each academic cycle. In 2008–2009, when we were assembling the machine, the department had eighteen full-time faculty members, so our machine could produce 360 scored samples. The primary goal of the methodology is to compare multiple evaluations. With eighteen readers scoring twenty samples, each reader could be paired with every other reader on at least one sample paper if we were content to limit "multiple" to two readers per paper. That limit permitted the maximum number of raw samples: 180. The PA was entrusted to collect complete sets of every paper submitted in every assessed course (and empowered to bring down the wrath of authority upon the noncompliant), and then randomly to sample proportionately from each course to get the finite number of raw samples that the assessment machine could process.

The PA's role became even more prominent as the collector and superintendent of data, some of which had to be treated as confidential. A prevailing assessment principle at LTU is that we are engaged in the assessment of student work, not of faculty performance. Assessment data must have no bearing on promotion, tenure, raises, access to faculty resources, and the like. Yet the central hypothesis of the new writing assessment process depended upon comparing faculty members' grading patterns against each other. In recognition of the seriousness of these principled and institutional issues, various measures were built into the new writing assessment process to maintain anonymity. Some of these measures necessarily required the PA to have exclusive access to information that they must then translate into a neutral coding for the open data set. Some readers may think these measures excessive. But they were incorporated with an eye

toward achieving unanimous faculty buy-in to the process. That is to say, they were designed to satisfy those few faculty members most suspicious of the institutional motivations behind assessment procedures.

In the 2008–2009 academic cycle, Composition and the first semester of the philosophy/history sequence ("Foundations of the American Experience") were already on the docket for their triennial assessment. Instructors had been directed to collect copies of all papers submitted in their fall sections. We were thus presented with 365 sample papers drawn from these two courses to test our model. The PA simply went through the giant stacks of papers and selected every other one to arrive at 182, and then picked two out from the middle. The PA redacted any information on the papers that identified the student, section, or instructor. He then made copies of the samples, reassembled each, and encoded each with a unique serial sequence identifying its provenance. A "scores" sticker was attached, containing nine boxes with labels corresponding to the nine fields of evaluation on the rubric, for the readers to enter their scores (see fig. 2).


F08/213 02 P2 0B016

[redacted]

Hinduism and Islam

There are many religions throughout the world. Religions have been passed on from generation to generation. Most people have a religion, while others don't believe that there is a specific religion that is right or wrong. Some people believe there are many gods, while others believe there is only one god. There are religions that are monotheistic, polytheistic, or even Atheist. Atheists meaning they don't believe in any god. In the book called, Bhagavad-Gita, the religion they follow is part of the Mahabharata of the Hindu religion. The main point of the story was to live your life to the fullest, to follow whatever you want, whoever you want, and to

SCORES	
A.Thes	
B.Dev	
C.Cou	
D.Com	
E.Rela	
F.Cita	
G.Sty	
H.Gra	
I.Para	

FIG 2.

Once the raw samples were encoded, they were then sorted into eighteen stacks of twenty papers, ensuring that each stack was paired with every other stack on at least one sample. The PA then went through each stack and added one additional number to the serial sequence on each paper: a "scorer number" that distinguished the samples in each stack. The eighteen stacks of papers, along with copies of the common rubric and an instruction sheet, were placed into plain manila envelopes, and the envelopes shuffled and placed in a bin. At the first faculty meeting of winter term, the bin was passed around the room and each faculty member chose an envelope, or "packet," from the bin. Faculty were required to initial a log marking their receipt of a packet (and again to initial their return of their packet), but this elaborate ritual ensured that no one (not even the PA) could know which faculty member was identified by which scorer number in the data. Each participant in the process could identify his or her own scorer number but no other participant's number. Faculty members were instructed to score their packet of papers and return the packet to the PA by the end of winter term. Figure 3 shows a portion of the data set that the PA assembled when all the assessment packets had been returned.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
1																															
2	Semester	Course	Section	P1/P2	Lot #		Copy	Scorer #	Thesis	Development	Counter-arguments	Course material	Evidence-claims	Citations	Style / concision	Grammar / syntax	Paragraphs	Total Score		Copy	Scorer #	Thesis	Development	Counter-arguments	Course material	Evidence-claims	Citations	Style / concision	Grammar / syntax	Paragraphs	Total Score
3	F08	1103	3	P1	1		A	21	4	4	3	3	3	2	2	4	2	27		B	13	4	7	0	3	7	0	4	4	7	36
4	F08	1103	3	P1	2		A	23	3	3	1	4	5	2	4	2	3	27		B	22	11	9	7	7	8	8	4	5	7	66
5	F08	1103	3	P1	3		A	25	7	8	0	8	5	9	6	8	11	62		B	24	5	5	0	5	5	2	5	5	5	37
6	F08	1103	3	P1	4		A	12	0	2	3	2	1	0	0	4	0	12		B	14	7	7	0	6	6	7	6	5	6	50
7	F08	1103	3	P1	5		A	28	5	6	0	4	5	9	9	10	6	54		B	11	5	7	0	3	8	7	5	5	6	46
8	F08	1103	3	P1	6		A	26	8	6	2	9	7	5	9	8	8	62		B	27	9	10	0	8	8	5	5	5	8	58
9	F08	1103	3	P1	7		A	19	4	5	0	5	5	0	6	4	7	36		B	20	6	5	0	6	5	7	4	7	6	46
10	F08	1103	3	P1	8		A	17	9	10	0	11	0	0	10	10	9	59		B	18	7	5	0	2	3	2	5	7	4	35
11	F08	1103	3	P1	9		A	15	7	10	0	9	8	0	6	5	9	54		B	16	11	11	0	9	8	0	10	11	11	71
12	F08	1103	3	P1	10		A	14	10	10	0	9	9	7	10	10	10	75		B	13	4	10	0	5	8	11	0	11	11	60
13	F08	1103	3	P1	11		A	12	0	0	0	0	2	2	0	4	0	8		B	21	2	2	2	2	2	2	2	2	2	18
14	F08	1103	3	P1	12		A	11	10	10	5	6	7	0	8	4	6	56		B	22	8	7	5	9	9	0	8	8	10	64
15	F08	1103	3	P1	13		A	28	7	7	0	6	6	6	9	11	8	60		B	23	3	3	2	3	4	8	6	8	3	40
16	F08	1103	3	P1	14		A	24	8	8	0	5	5	5	8	8	5	52		B	27	8	7	5	4	5	5	4	5	6	49

Sheet1 Sheet2 Sheet3

FIG 3. Data set constructed from assessment results

Analyzing the Data I: Reliability of the Instrument

Analysis of this data proceeded with an interest in addressing two broad types of questions. The first type concerned the statistical validity of the common rubric and consistency of scoring among pairs of readers. If the answers to these questions resulted in confidence in our method, another set of inquiries could then sort latitudinal and longitudinal trends in reported student-skill levels. As the process unfolded, it became clear that in this, as in every other element of the project, the cultural dimension of communicating the results to diverse audiences would require accommodation. We were committed to subjecting the data to a level of statistical rigor that would impress our natural science colleagues across the university. But it very soon became evident that the kinds of logarithms, standard deviations, *P*s, *kappas* and so on that SPSS provided meant very little to the great majority of the humanities faculty. The processing of the data into charts thus followed two parallel tracks: one set for a statistically grounded audience, and another set for an audience who tends to look at charts more like pictures than formulas. (And it is worth emphasizing at this point that the author of this article is decidedly of the latter disposition.)

An accepted statistical method for measuring the validity of instruments for recording observational data, such as rubrics, is to perform a paired-samples *t* test upon the results. In laymen's terms, this logarithm evaluates the likelihood that multiple observers are basing their ratings on something other than the thing being tested (in our case, the scale set in the rubric.) The results, shown in figure 4, indicate an exceptional degree of consistency, or concordance, among the scorers. Statistically significant variation is recognized when the negative standard deviation of one observational set exceeds the positive standard deviation of the paired observational set (when the lower deviation bar of the taller column does not overlap the upper deviation bar of the shorter column). In only one of nine fields of comparison does the *t* test indicate significant variation ("Citations"), and even here the results barely reach that standard. In several other fields, the deviation bars are almost parallel. These results are a very strong endorsement of the methodology adopted.

However, these results did not communicate very much information to the majority of the humanities faculty. That our natural science colleagues

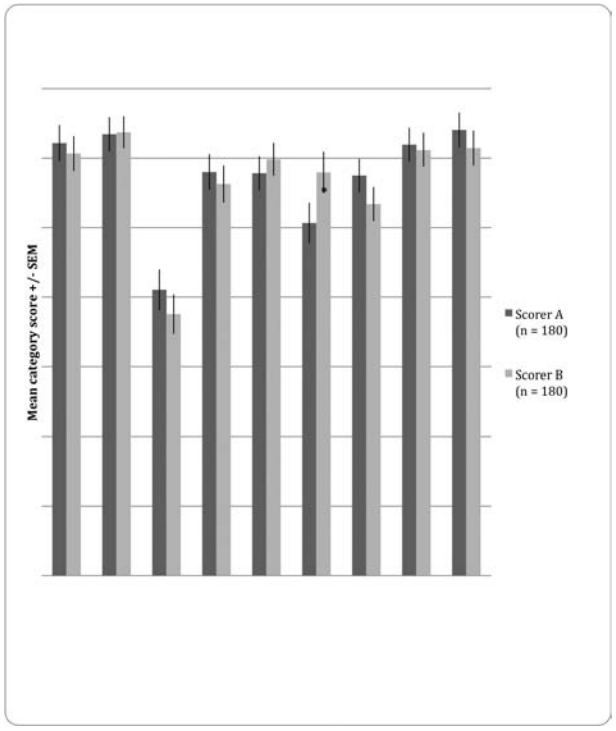


FIG 4. Concordance between scorers: Paired samples t-test
* P < .05

would be obligated to acknowledge the validity of the results was heartening, but few of us had even the vaguest notion of how a standard deviation was calculated or of the significance of overlapping deviation bars. The validity of the method seemed to many of us to rest upon a fairly simple question: when any two readers looked at the same sample text, to what extent did they give similar scores? Figure 5 illustrates one attempt to get at that question. For each field of evaluation, figure 5 shows the distribution of variations among the 180 paired scores: how often each reader gave scores that were the same or one point different, two to three points different, four to five points different, and so on.

Illustrated this way, the data provide a more detailed view (at least for those ignorant of the conceptual relationships invoked in statistical logarithms) of the degree of consensus among the readers in the application of the rubric. But because this illustration follows no accepted statistical

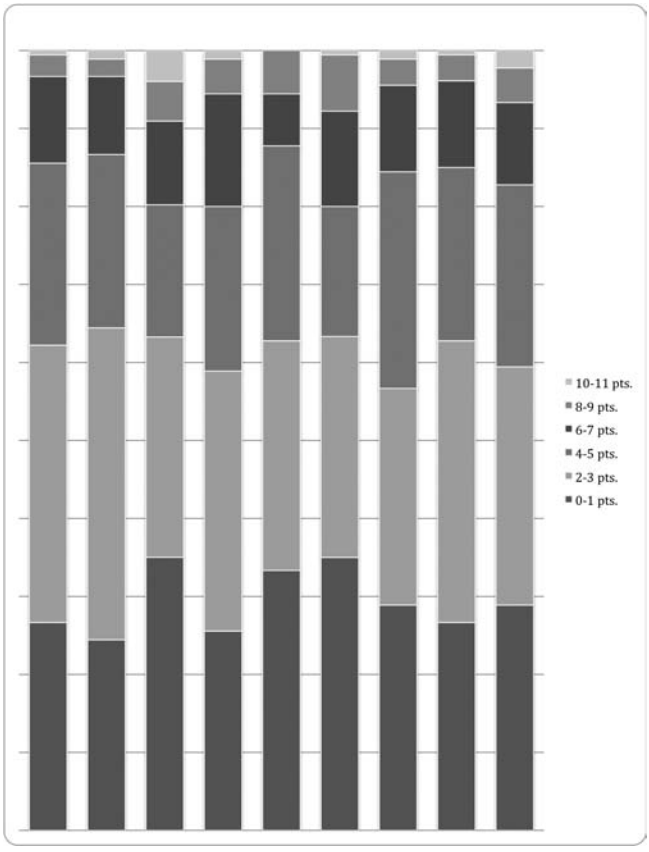


FIG 5. Paired scorers' variation by category

logarithm, it is difficult to establish benchmarks for degrees of success. If we note that two readers who scored within three points of each other were in the same letter-grade ballpark, we could then assert that just about 60% of all the pairs analyzed in figure 5 met that standard. This remains arbitrary as it indicates nothing of whether 60% is an acceptable rate for meeting that standard. In this as in several other issues with benchmarking that emerged, we will not be able to establish defensible markers until we have implemented the process several years in a row. At that point, we will be able to look back longitudinally and see norms.

In order to identify more clearly the sources of the variation found in figure 5, figure 6 calculates and compares a total variation number for each scorer. This number is determined by summing the variations in all

of a given scorer's entries compared to their paired entries.[3] Once again, there is no statistically accepted way of determining significant variation in this context and we will have to wait for longitudinal comparisons to see whether the first-year's results are empirically "normal." In the meantime, another offhand metric could be applied based on an average variation of 1 in each paired entry. If a scorer averaged +1 across all 180 of their entries, their net total variation would be +180. We might then think of 180-point increments in figure 6 as corresponding to one-point increments on the rubric. By this metric, figure 6 shows that ten of eighteen scorers' total variation numbers fell within an average one-point difference with their paired scorer on the rubric, and sixteen of eighteen fell within the two-point range.

This illustration of the data strongly indicates two features of the range of perspectives among the reader pool. First, it is evident from these data that a handful of scorers in the pool accounted for an inordinate amount of the variation found in the cumulative data set. Second, the pattern for at least half the reader pool was that their scores registered high compared to some other readers, low compared to others, high in some fields of evaluation, low in others. (Additional charts, not included here, break down each scorer's variation by field and show ranges of variation.) In this light, our assessment machine appeared to be processing the myriad differences in perspectives among the faculty into something like mathematically determinable norms (even if we would not be able to calculate some of those norms for some years yet).

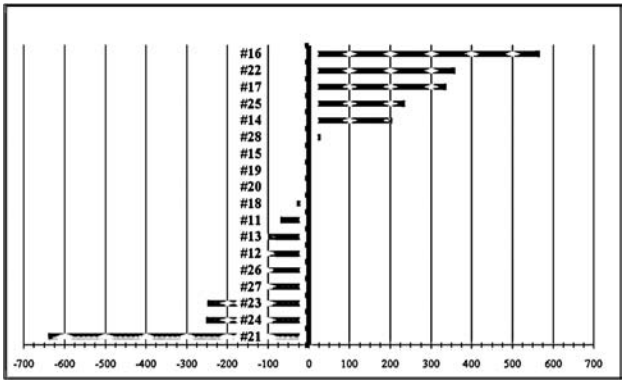


FIG 6. Total score variation by scorer

Analyzing the Data II: Student Performance

Each of these approaches to demonstrating the validity of our process reinforced our confidence that we had "good numbers" with which to work. Turning to what those numbers could tell us about our students' writing skills again required parallel series of charts, one series for the statisticians and another for the humanists. Means comparisons permit the calculation of standard deviations and calibrating of significant variation. Figures 7 and 8 are two examples from a series of means comparisons performed, and they illustrate the two broad questions we sought to answer. Was there significant variation among the sections? And, was there significant variation from the first paper set (P1) to the last paper set (P2)?

Figure 7 addresses this first question in relation to three sections of one of the courses sampled. The comparison indicates that two of the three sections produced papers very consistent with each other, while the third section's papers scored substantially lower in every field of evaluation. Two inferences are reasonably made from this disparity. There may have been a difference in the pedagogical effectiveness in the sections being sampled, in which case the anonymity of the process proves crucial, for there is little

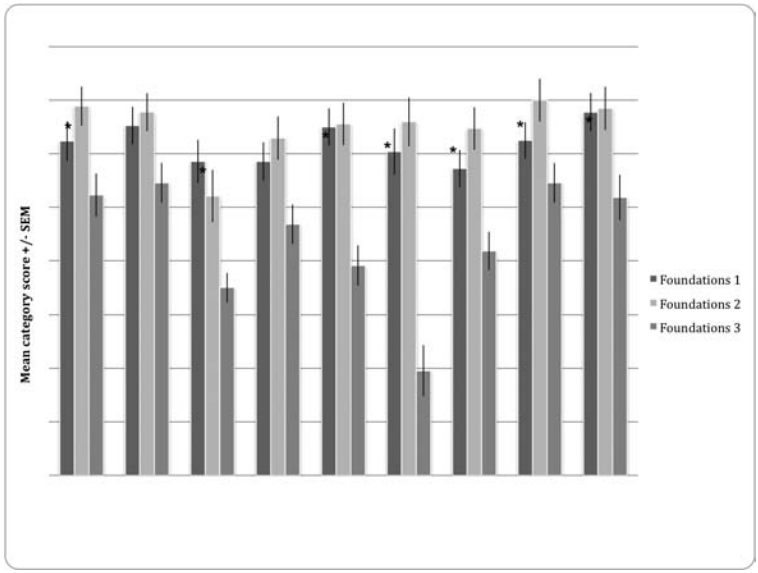


FIG 7. Mean scores on both papers: Foundations sections
* P < .05 Difference

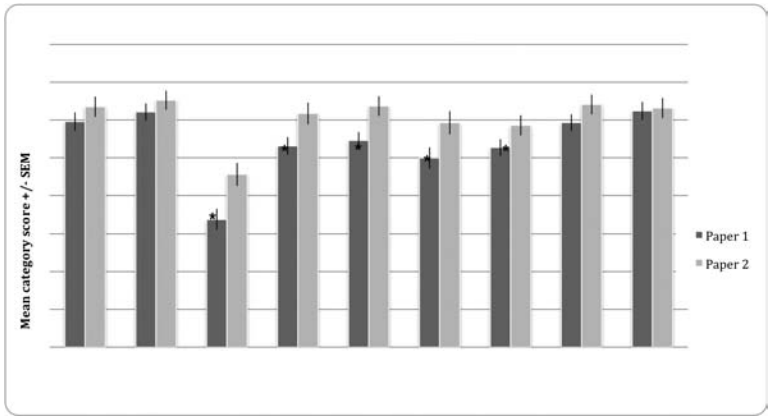


FIG 8. P1 v P2 means comparison: All samples
* P < .05

prospect of significant faculty buy-in to the process if it was understood that the published results would identify individual faculty members. Another reasonable inference is that section 3, or the samples taken, contained a cohort with particularly weak writing skills. Data illustrated below reinforce this second inference.

Figure 8 is one example of a series of charts that addressed the second broad question regarding student performance. Figure 8 illustrates a means comparison of first and last papers for all the samples collected. These data show statistically significant variation from first to last papers in five of the nine fields of evaluation, each one varying higher in the latter set. In all nine fields of evaluation, the real mean of the second set was at least nominally higher than the first set. These results are a very positive, if modest, affirmation of pedagogical effectiveness. A semester of work with our faculty appeared to have a net positive gain for our students. That this progress registered in relatively consistent differences across all nine fields of evaluation further affirms that the common rubric is a fair barometer of the breadth of the faculty's curricular goals.

The real means recorded in figures 7 and 8 do not overly impress one with the caliber of our students' writing abilities. Most of the means fall in the C to D+ range. The particular courses from which the samples were drawn tend to be populated with first-year students (Composition almost entirely so, Foundations more mixed with second-year students). Additionally, the samples were drawn from fall-term sections. The humanities faculty members are familiar with the deficiencies of our first-year students,

so these means are not greatly surprising. These means will provide a baseline against which we may compare samples assessed from later Core courses. Low means in the early Core courses may prove an advantage. One would hope that there is nowhere to go but up, and the longitudinal differences ought to be fairly accountable to the impact of the writing curriculum in the Core.

The fundamental difficulty with the means comparisons in figures 7 and 8 for faculty who regularly teach writing is that we see very few mean students in our classrooms. The columns in those charts do not *look* like the lots of twenty papers we receive in each of our sections two or three times each semester. Figures 9 and 10 are two examples from a parallel series on student performance that expresses proportions of scores in four point categories. Figure 9 compares the proportions of these point categories in each of the courses sampled and the cumulative body of samples in each of the nine fields of evaluation. These proportions *look* much more like the collections of students we typically find in our classrooms, and permit the faculty within each cohort to see how their students are distributed

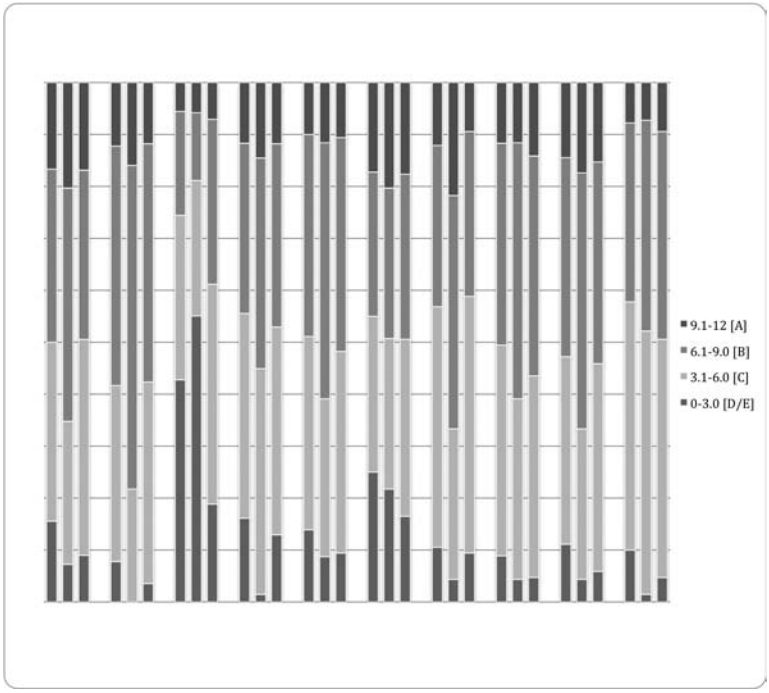


FIG 9. Proportions of scores on both papers

compared to a norm for the entire Core Curriculum. The results are mixed in this first year's sample. Papers from the two courses scored in varying patterns, with no apparent trends. The significance of these proportions may become evident in succeeding years, as we are able to do longitudinal comparisons.

Figure 10 is one example of a more direct attempt to measure differences between the proportions. Figure 10 compares the proportions of samples in four total-score categories for P1 and P2 samples for each of the sections sampled. Again the results are mixed, but at least one indicator of progress from P1 to P2 samples can be seen in the highest score category. Four of the six sections had no "A" scores among the P1 samples, and then had 10–20% of their P2 samples score in that range. Of the two remaining sections, one increased the number of highest-scoring papers, and the other had more than 30% of its P2 samples score in the "B" range where it had had none of its P1 samples score that high. Every section made at least nominal progress at the top of the scale. The patterns are more mixed at the bottom of the scale. This difference may be the starting point for an insight into our curriculum: is our writing pedagogy better suited for students who arrive at LTU ready to take off into college-level writing than it is for students who arrive ill-prepared?

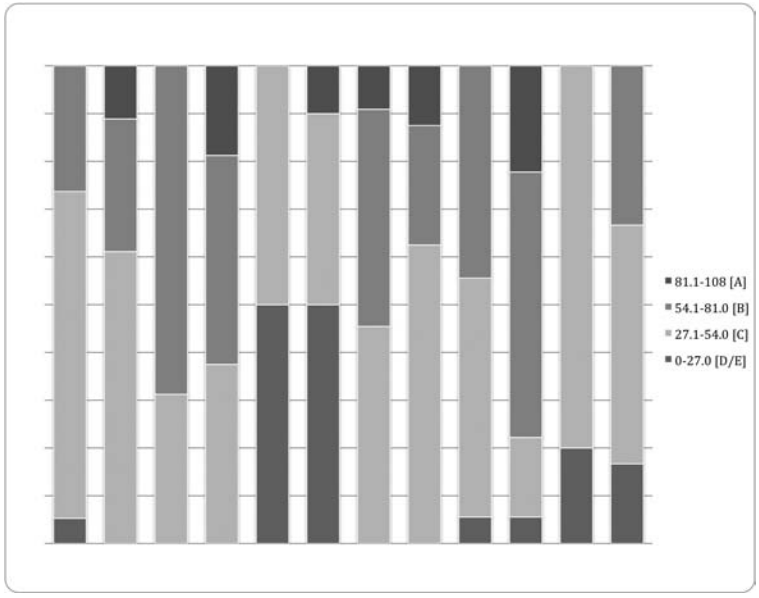


FIG 10. Proportions of total scores P1 v P2

Figures 4–10 illustrate the two broad types of questions posed to the raw assessment data, and the two broad audiences for whom answers had to be interpreted. As indicated, a third type of inquiry—longitudinal analysis—will become possible as we incorporate successive assessment cycles into our data set. The ultimate goal is to build a complete tracking record for the six Core Curriculum courses in a single data form, refreshed on a three-year cycle and maintained by a sustainable labor commitment from faculty.

Dividends on the Humanities Faculty's Investment

Some of this assessment machine's utilities have already been realized. Other utilities are anticipated. The process of developing the machine was highly constructive in terms of departmental culture and the faculty's esprit-de-corps. The discussions upon which the machine was built were pedagogically substantive and constituted one of the most thorough internal reviews of curricular standards conducted by the department in many years.

The assessment machine has also proven effective in restoring the Humanities Department's authority over writing assessment across the university after the debacle of the report on undergraduate seniors' writing skills. In the spring following that report, the PA toured the faculty meetings of the other colleges with a presentation on the assessment process the humanities faculty had developed. Several features of that process impressed our natural science colleagues. First, it was evident that the humanities faculty had taken the problem of writing assessment very seriously and invested an inordinate amount of time in working out the methodological problems raised in the debates from the prior fall. Second, it was evident that the humanities faculty had made a strong effort to meet empiricism upon the natural scientists' ground. We did not retreat behind the claim that writing is so inherently subjective a process it cannot be measured with numbers. Instead, we developed a quantitative system that we thought incorporated and accounted for the complexity of evaluative factors. Finally, the assessment machine's results were varied, showing a broad distribution of skill levels among our students and a modest cumulative impact of the writing curriculum. If our results had been as cartoonishly positive as the prior report's results were negative, they would have elicited a similar skepticism. From a kind of unaccountable

embarrassment, writing assessment has become one of the department's institutional assets. The university's provost has indicated her desire that a presentation on the assessment process be a featured item on the itinerary of our HLC accreditation visitors in the fall of 2010.

The assessment machine's data output has also become the basis for curricular and institutional reforms within the department. The humanities faculty was generally mortified to discover that "citations" proved to be one of our students' weakest skills and one of the fields with the greatest variation among scorers. We had considered citations one of the fundamental elements of our curriculum. The first year's assessment results led to discussions and a renewed commitment to emphasizing the necessity for students to acknowledge their sources in the correct citation format.

The assessment machine has provided the department with a quantitative foundation upon which to make more substantive proposals for institutional changes. For example, a long-standing defense by LTU's humanities faculty to the charge of ineffectiveness in our writing pedagogy has been the observation that more substantial progress cannot be expected with the university's 4/4 teaching load. Given the skill profiles of the student body and the dearth of writing pedagogy outside of the humanities Core Curriculum, it is simply unrealistic to expect the humanities faculty to provide the kind of intensive, individualized instruction necessary to bring students to a level of writing proficiency that will promote their career prospects when those faculty are working with 70–80 students per term in four sections of two or three courses. The assessment machine has permitted the department to formulate that customary complaint into a positive, testable hypothesis and proposal for university resources. The department has requested funding to permit two faculty members' teaching loads to be reduced to a 3/3 (either the same two members for several years or rotating each year). The department will then sample the writing from those faculty members' students and process them (anonymously) through the assessment machine alongside samples from sections whose instructors are teaching a 4/4 load to see if there are measureable differences in pedagogical impact. The provost has been highly receptive to the hypothesis and means of measurement, but thus far has pled ongoing budget constraints, given the prevailing economic climate. Whether or not the project is funded, the proposal—and its undergirding by an assessment process that meets empirical standards—has shifted the weight of institutional accountability. We are now prepared to measure on the natural scientists'

terms the impact of changes we humanists know would be most efficacious for the subjective skills we are trying to develop in our students.

In these and other ways, the department's writing assessment process has already repaid the investment of time and energy into its development. Future dividends will depend upon the faculty's management of this new institutional asset.

Notes

1. The humanities component of the Core Curriculum consists of six specific courses and a third/fourth-year elective: one semester in composition and one in communication, a two-semester series in Western art and literature, a two-semester series in Western history and philosophy, and a junior/senior-level topical seminar.

2. The university's assessment committee collected sample papers from third- and fourth-year courses from across the university in the spring of 2008. Over the summer, a panel consisting of one faculty member from each of the colleges evaluated fifty of the sampled papers, selected proportionately to reflect enrollment figures in each of the colleges.

3. For example, on a given sample paper Scorer 11's marks across the fields of evaluation are +1, +3, 0, -1, -2, 0, +4, +2, and +2 compared to Scorer 12's marks. Scorer 11's sum variation would be +9 on this sample (Scorer 12's sum variation would be -9.) The total of Scorer 11's variations thus calculated for each of the twenty papers scored constitutes the total variation number. The theoretical limits of this scale thus run from -2160 to +2160 (a twelve-point variation on every one of 180 paired-samples).