Assessing Critical Thinking through Student Writing:

An Academic Model of Accountability

-----

Richard Fliegel and John Holland

USC College

On May 16th, 2007, The American Association of Colleges and Universities sent out an “alert,” anticipating changes in the federal regulations that govern accrediting agencies. Despite the contentiousness of Secretary Margaret Spellings’ Commission on the Future of Higher Education, the AAC&U email warned, the U.S. Department of Education is moving forward on a recommendation to institute “a very small set of standardized tests administered outside of a student’s regular course-taking.”¹ In the AAC&U’s journal, *Liberal Education*, Judith Eaton predicts this “would seriously erode the successful self-regulatory enterprise of the past hundred years” and threaten “the longstanding leadership of the academy in determining academic standards and judging academic quality.”²

Eaton views the initiative as an attempt to seize control by radically expanding federal authority over the content of higher education. But the Secretary’s Action Plan, to “convene members of the accreditation community to recommend changes to the standards for recognition that will place a greater emphasis on results,” reflects rather an investor’s impatience with the academic community’s reluctance to meet demands for accountability with evidence of student learning. “We expect detailed information about

¹ Email citing a September 2006 statement of the AAC&U board on the Spellings Commission Report, at [http://www.aacu.org/about/statements/Spellings9_26_06.cfm](http://www.aacu.org/about/statements/Spellings9_26_06.cfm).

² Forthcoming.
nearly every purchase or investment we make, from choosing a doctor to buying a car or home,” she told participants at a summit in March 2007. “Yet when it comes to higher education, we’ve invested hundreds of billions in taxpayer money and just hoped for the best. Colleges and universities must be more transparent about cost, graduation rates, and learning outcomes, so that students and parents can make more informed choices.”

An increased emphasis on learning outcomes is already required when institutions of higher education are reviewed for accreditation. Ralph Wolff, executive director of the WASC accrediting commission, writes, “Accreditation as an agent of accountability has shifted significantly from one of process (that is, everyone goes through a review) to one of outcomes, with an increasing focus on student success and learning.” Similar demands for data about learning outcomes have been coming from other stakeholders for years. "State policymakers are rightly asking what policies will enhance the effectiveness and productivity of their higher education systems," Richard Hersh and Roger Benjamin told the National Governors' Association in 2001. "The logical and the only real candidate for accomplishing this task is to look at the consequences of new policies and program improvements on the outcomes of student learning. Current assessment data, however, are not satisfactory."

---

3 Secretary’s remarks at a March 22, 2007 summit on “A Test of Leadership: Committing to Advanced Postsecondary Education for All Americans.”


It is tempting for government agencies to attribute the lack of acceptable evidence to the recalcitrance of educational institutions whose reputations and financial interests may be challenged by quantitative comparisons of outcomes. Robert Zemsky and his colleagues have observed in dismay that parents and students seem to choose colleges by the “competitive advantage” presumed to follow a school’s prestige, rather than by the evidence of student learning outcomes at an institution, even when the latter is available.\(^6\) The most prestigious schools have not led the call for articulated learning objectives and published data concerning value-added student accomplishments. But the challenges involved in obtaining good data, especially of the higher-order skills most essential to a college education, prevail even where the will is not lacking. Institutions of higher education often face a choice between assessing what is amenable to assessment or trying to evaluate what is central to our enterprise but much more elusive. Because of the tendency to bog down in debates over definition and evaluative strategies, Karl Shilling has called critical thinking “the Vietnam of assessment.”

The centrality of critical thinking to American college curricula is apparent from any survey of the mission statements used to justify the academic requirements of undergraduates. Harvard College encourages its students "to respect ideas and their free expression, and to rejoice in discovery and in critical thought."\(^7\) Yale’s distributional requirements provide competence in "the languages of thought," because "such competence gives you the tools to think critically and analytically, and to enlarge your


imagination." As President of Amherst, Tom Gerety summed up the intentions of his own best college teachers when he told his graduating students, "The bigger, more enduring question, the one you will have to ask yourself as long as you live is this: Are my thoughts and stances my own? Do I deserve the respect of these teachers, deserve it because I am what they deeply want me to be -- not a clone of theirs or a convert but a free, independent thinker, an adult with the courage and tenacity to come to my own convictions?"

"Critical thinking" is widely seen as the defining element of American higher education, the requirement for informed participation in a democratic society. It is the core objective of a liberal education, a goal motivating the restructuring of college curricula and the development of new learning-centered pedagogies. A great deal of thoughtful work has been done identifying the constituent skills of critical thinking (e.g., Kurfiss; King and Kitchener; Jones, Dougherty, Fantaske, and Huffman; Wolcott) and issues related to its assessment (White; Nosich; and Erwin). Yet, despite a

---

9 Thomas Gerety, Convocation address, Amherst College, 1997.
shared perception of its value to our citizenry and workforce and its centrality to the mission of higher education, we do not have an acceptable measure of success in teaching critical thinking.

Why is it so difficult to evaluate a school's success in teaching their students to think critically? The sharpest lines of disagreement fall between the operating principles of assessment experts and the culture of colleges and universities. At a meeting of the AAC&U in Virginia, T. Dary Erwin stressed the importance of choosing an appropriate instrument but warned that "you might not find one." While articulating goals that are amenable to assessment, Professor Erwin characterized as "loosey-goosey" a set of goals such as "education for lifelong learning," near to the hearts of research faculty and the administrators who rise from their ranks. Psychometricians such as Steven Klein at Rand challenge the reliability and validity of faculty assessments of their own students’ work. Faculty have always assessed their student learning for the purpose of assigning a grade, of course, but if those assessments were to be considered data in evaluating the success of a program, the faculty would in effect be assessing their own performance. Measures of student engagement, such as the National Survey of Student Engagement developed by George Kuh at Indiana University, provide valuable insights, but student opinions are no more reliable than faculty opinions as indicators of academic performance.

---


On the other hand, from the perspective of faculty who are not experts at learning assessment, goals articulated as "competencies" reduce interpretive critical thinking to a decontextualized series of more-easily-measured skills. They exhibit little enthusiasm for fitting their classroom practice with input and output measures on the model of a social scientific research experiment. “Critical thinking” on these instruments is abstracted beyond recognition; their format does not allow for the complexity of thought intrinsic to natural writing; administering them is invasive of class time and spirit; and "those tests don't measure what I teach."

The National Center for Education Statistics evaluates standardized instruments for assessing critical thinking, including the Academic Profile, the Collegiate Assessment of Academic Proficiency, the CAAP Critical Thinking Assessment Battery, the California Critical Thinking Disposition Inventory, the California Critical Thinking Skills Set, the Cornell Critical Thinking Test, the College Outcomes Measures Program, the ETS tasks in critical thinking, the Measure of Intellectual Development, the Problem Solving Inventory, the Reflective Judgment Inventory, and the Watson Glaser Critical Thinking Appraisal. Some of these instruments have been around for a long time: the Cornell test, for example, was first published in 1971, and the Watson Glaser in 1980. Most use a fill-in-the-bubble format; some elicit writing samples as part of the test material. Roger Benjamin has developed an in-basket exercise that elicits sophisticated thinking, but what undergraduate fails to see an in-basket exercise as outside the scope of his or her normal class requirements?

17 Website at nces.ed.gov.
To some extent, “embedded” assessment methods enable faculty to address the shortcomings of standardized tests by incorporating measures into their own classes. This approach works best for individual faculty committed to assessment, or small groups of like-minded faculty, to ensure that evaluation is related to the objectives of a course. Its utility diminishes, however, in assessing larger or less uniform academic programs, when the metaphor shows its corners. “Embedding” implies incorporating or concealing some element extrinsic to the classroom practice, and when that proves to be the case, faculty resent it and students do not attempt to do their best work.

Academic programs intended to help students become generally well-educated people who can convey their thoughts effectively in writing have been particularly hard pressed to show that students are learning what these programs have been funded to teach. What is needed is a system that reflects the values of academic culture but allows us to assess the changes in critical thinking students demonstrate over the time they are engaged in their general education classes.

Colleges and universities need to design our own assessment methods to forestall the imposition on accrediting agencies of federal guidelines that require inappropriate assessment tools at institutions of higher education. These campus-defined methods must respond to the demands for public accountability and yet measure what we believe is most important for our students to learn. They will need to be more effective than those which are likely to be developed by (or attractive to) the U.S. Department of Education, evident in the system of standards and tests at the elementary school level.

To elicit from students their true levels of capability, we will need assessment methods that students perceive as part of their regular curriculum. This will only be true when faculty take ownership of the process and believe that the skills being evaluated are those they consider aligned with their own learning objectives – not the outcomes most easily articulated but the intellectual capabilities without which no student can be judged an generally well educated person, capable of the critical distinctions essential for membership in a democratic society. This kind of assessment is the most difficult to do by standardized tests, but is accessible through the academy’s historically most valued method of assessment: student writing.

Assessing Critical Thinking through Student Writing

Stage One

The approach taken at the University of Southern California (USC) is based upon two considerations. First, we believe that faculty are the best judges of their students’ academic performance and can be relied upon to make valid assessments, if a context can be structured that diminishes their personal investment in the outcome attributed to any particular group of students. Second, we are situated in a curriculum that requires all students to take one writing class at the freshman level and another at the junior or senior level. This curricular structure enables the growth of writing skill to take advantage of developmental changes experienced by students as they progress further from high school and closer to the graduate, professional, or working world in which writing skills are evidently valued. This structure also allowed us to elicit samples of student writing in relatively controlled conditions, so that the critical thinking content of essays written by
entering students could be compared to those by students who have spent some time in our program. Our goal was to train a cohort of teaching faculty to evaluate the critical thinking of essays written by students in both classes, without knowledge of the level of the students whose papers are being read. If we could demonstrate a significant increase in the scores of students in the later classes, in essays assessed specifically for evidence of critical thinking, we could demonstrate in the first stage of our project that this method of assessment can quantify the value added by their experience at USC.\textsuperscript{19}

We have not claimed, nor hope to claim, that the increase in critical thinking skill, as assessed by our scorers, is attributable to any particular college class or set of classes. A wide variety of experience contributes to the development of critical thinking, inside and outside of the classroom, and our project is not designed to differentiate, for example, between students of the same level who have or have not taken a particular writing class. This is a global assessment, focused on a key learning objective of the General Education Program but one that is also part of our university’s central mission.

In the first stage of our project, we selected a rubric for scoring critical thinking content, trained a cohort of 15 readers, and evaluated the content of 475 student essays, written in response to one of two assigned prompts in a timed-writing situation. These conditions were chosen to reduce the number of complications that had to be addressed in creating a valid and reliable process. Our eventual goal is to use essays assigned by their faculty under more natural conditions, but for the first reading we controlled as many variables as possible.

\textsuperscript{19} The USC General Education Critical Thinking Project was funded by a grant from the USC Provost’s Office and the College of Letters, Arts & Sciences
Two prompts were used so that a longitudinal study would be possible at a later time, when the students who had written the freshman papers were juniors or seniors, who might then be asked to respond to a prompt they had not seen before. For the first reading, the papers were identified by nondescript numbers, divided by prompt, but mixed together so that no reader could determine whether any individual paper had come from the freshman class, Writing 140, Writing and Critical Reasoning, or the upper-division class, Writing 340, Advanced Writing. The readers were trained in a process of “socialization” that has a long history in writing programs and at USC in particular.

An “anchor set” of eight essays was identified by three Chief Readers, who read through a large number of essays from both the 100-level and 300-level classes and chose four essays that demonstrated what each reader considered to be sharply differentiated evidence of critical thinking, and another four essays chosen to provoke discussion points among the socializing readers.

To guide the training and scoring we used the Holistic Critical Thinking Rubric designed by Peter and Noreen Facione. Peter Facione is currently Provost of Loyola University in Chicago. He has had a long career in the scholarship of critical thinking, as a Professor of Philosophy, as an author on the subject, and as a creator of a standardized test of critical thinking. The Facione scale has four points, an upper and lower half, differentiated by evidence of critical thinking. Among his publications Professor Facione includes a summary of the conclusions of a generation of scholars writing about critical

---

thinking, and the Facione rubric reflects those collective conclusions. We did raise questions about the lack of attention paid by the scale to insight or originality of thought, but those did not prove to be issues when the essays were read.

The results of our first reading are summarized on the chart below, which graphs the upper-half versus lower-half differentiation resulting from the first reading. Each essay was read at least three times by different readers; “gaps” of two-point difference, such as a score of 4-4-2, were read again by chief readers. Finer distinctions, within each half of the scale, produced a more complex result, including in one case unusually high scores for the students of a particular instructor. However, the first judgment readers were trained to make marked the upper-half, lower-half distinction, and the distribution of scores on that measure was encouraging:

---

The results indicated that fewer than half (45.3 percent) of the papers written by students in the freshman writing class (140) were judged by trained readers to fall into the upper half of the scale, while 58.2 percent of all papers written by junior or seniors fell into the upper half of the scale. When transfer students are removed from the Advanced Writing (340) cohort, so that only students who completed both writing classes at USC remain, the total rises to 65.5 percent of the papers in the upper half of the Facione scale; and when Advanced Writing students in USC College are differentiated from those in our professional schools, the total rises to 73.8 percent.

We would argue that these results can be used to quantify the difference in critical thinking skills of USC students at various points in their college careers, and to validate the process of assessing critical thinking through samples of student writing. The random chance of this pattern emerging in the absence of those two factors is highly unlikely.

**Stage Two**

To address the faculty concern over the validity of this method, expressed as, “Those tests don’t measure what I teach,” nine experienced faculty members agreed to rank-order the anchor set of eight essays from the Stage One reading, using whatever criteria they normally employed in practice. These faculty included participants from the Departments of English, History, Linguistics, Philosophy, Political Science, Psychology, Slavic Languages and Literatures, and Spanish and Portuguese. The group included a former Dean of Undergraduate Studies, some Department Chairs, and recipients of every important teaching award bestowed by USC.
Our expectation was that their rankings would disagree with one another, but that the relative weight assigned to papers that demonstrated more or less critical thinking would approximate the judgments of our socialized cohort of Stage One scorers. We did not assume the faculty rankings would be more or less valid measures than the holistic scores awarded by our readers in Stage One -- simply that the faculty in Stage Two would rank the essays using the same sense of “critical thinking” they bring to their classes. If the holistic scores in Stage One reflected judgments similar to the faculty rankings, we could convincingly argue that we were in fact assessing the same set of skills they had in mind when they taught “critical thinking.”

These faculty were asked not to give each essay a grade, but simply to indicate which essays demonstrated greater skill in thinking critically. No other training or socialization was provided, and participants did not know who among their colleagues were participating. Comments were invited, but no definitions of critical thinking, explicit ranking criteria, or explanations for their decisions were requested. The effort required to arrive at a consensual definition of critical thinking has sometimes exhausted the time and energy available for its assessment.

As we predicted, individual faculty rankings differed from one another and from the judgments of our Stage One scorers: no two faculty produced the same rank order. However, when their rankings were assigned point values, averaged, and ranked by their means, the resulting order mirrored the order obtained in the Stage One reading by our cohort of socialized scorers. Once again, distinctions within the upper and lower halves of the scale reflected a considerable variety of individual preferences, but the upper-versus lower-half differentiation aligned: the same four essays were scored in the upper
half of the scale by the socialized scorers in Stage One of our project and the experienced faculty readers in Stage Two.

The first column shows the rankings of the essays in the “anchor set” used to socialize readers in Stage One: Essay 4 was judged the strongest, followed by Essays 6, 8, and then 1. These same four essays were judged by the experienced faculty (identified in the chart above as Readers A through I) to fall in the upper half of their rankings, though in a different order (Essays 8, 4, 1, and 6.)

Recall that the first four essays of the anchor set were chosen as relatively clear examples of the points of the Facione rubric, to differentiate scores; the second four were chosen to raise debatable questions about essays that accomplished some of the rubric’s
criteria more successfully than others. Some of the readers in Stage One preferred Essay 3 to Essay 7, while others argued for the opposite order, and a similar split of opinion was recorded for Essays 2 and 5. Disagreements among the experienced faculty rankers in Stage Two echoed those of the socialized scorers in Stage One, concerning the relative value of critical qualities to be emphasized. However, when the first four essays are broken out of the anchor set, the sequence defined by Stage One scorers (4-1-3-2) was recreated by the mean rankings of the faculty working independently in Stage Two.

Once again, the strongest result differentiated the essays that fell into the upper half of the scale (Essays 4 and 1) from the essays that fell into the lower half of the scale (Essays 3 and 2). The finer distinction between the two upper-half papers once again reflected personal preferences. Both essays responded to a quotation from Robert Bellah, but Essay 4, which analyzes the war in Iraq, edged out Essay 1, which begins by citing Aristotle and then falls off in development. Faculty rankings were closely divided on the
relative merits of those two upper-half essays, but there was strong agreement among the faculty that those two essays demonstrated more evidence of critical thinking than either of Essays 2 or 3.

It is worth noting that the Stage Two faculty read these papers a full year later than the Stage One scorers, which may have influenced their assessments of the analysis in Essay 4. The author of Essay 4 argued that the war in Iraq demonstrated the utility of Robert Bellah’s observation. However, the war in Iraq and public discussion of the war changed a great deal over that time, altering the intellectual context in which Essay 4 was read. A tenth reader, whose rankings were done weeks after the first nine participants, relegated Essay 4 to the lower half of the paper set, citing its “extended and detailed (and commonplace) descriptions of failures in the Iraq war.” A year earlier, the relevance of the writing prompt to the Iraq war had impressed scorers as less of a commonplace. Over time, the social context of the assessment seems to have been a factor in evaluating the originality and consequently the critical insight of the author’s arguments.

Responsiveness to the prompt marked an important distinction. Essay 8 is the longest paper, based largely on examples from personal experience. Essay 4 offers an analysis of the United States’ policy in Iraq. Each responds differently to the prompt, which provided a quotation from Robert Bellah followed by a question comparing “short-term fixes” to “long-term solutions.” Essay 8 parses the components of Bellah’s observation; Essay 4 illustrates its utility in an analysis of the shortfalls of American foreign policy. Which sort of response should be privileged as demonstrating a greater degree of critical thinking?
In their ranking and comments, the faculty valued essays that responded most specifically to the language of the prompt, which the Facione rubric does not emphasize. Our intention is to develop an assessment process that can be used without reference or even knowledge of the prompt, so the readers in Stage One were instructed to be guided by the Facione rubric. The Stage Two faculty reflected their usual practice in grading student papers by ranking the most responsive papers most highly.

Nonetheless, on the basis of these results, we would argue that our method does in fact assess what experienced faculty members mean by "critical thinking," despite the differences among disciplines, even when no criteria have been collectively articulated. Through their practical judgments our faculty rankers and socialized scorers found common ground, without facing the obstacle of a universally acceptable definition of the term, “critical thinking.” From the comments of our participants, we believe a rubric could be articulated that would capture finer distinctions among essays more consistently than our experience has thus far allowed. All evaluations should be done at the same time, to avoid changes in the intellectual context that influence readers’ judgments about insight and originality of thought. Our procedure for choosing an anchor set and for socializing readers could doubtlessly be improved as well. But the method described above promises to be an effective means of assessing critical thinking, while respecting the complexity of faculty variation and the integrity of the classroom experience.
Stage Three

As the USC General Education Critical Thinking Project develops, we plan to make use of papers assigned in more natural class settings. Our second reading elicited essays once again from students in Writing 140 and 340, but instead of using the same prompts in timed-writing situations, we collected papers written in response to a variety of prompts assigned by the faculty member, under guidelines established by the USC Writing Committee and the Advanced Writing Programs in three Schools: the Marshall School of Business, the Viterbi School of Engineering, and the USC College of Letters, Arts & Sciences. We have just begun to look at these essays, and the questions raised by their differences in length, genre, and topic. Preliminary results from a reading of these essays suggest a distribution of scores by academic level similar to that obtained in the Stage One reading, but the data will need to be reviewed more carefully before any conclusions are drawn.

However, even before the latest results are digested, we have seen improvements in the undergraduate program at USC stimulated by the assessment method created for the General Education Critical Thinking Project.

As a result of Stage One, we have trained a cohort of writing instructors to think carefully about critical thinking, particularly as it is demonstrated in student writing. Though we did not ask them to do so, several of the faculty participants in Stage Two of the project could not resist explaining why they arrived at the rank-ordering they did, articulating functional definitions of critical thinking that have helped us understand differences linked to disciplinary perspectives. These criteria will be used to draft a new rubric designed to increase reliability on the finer points of the holistic scale.
The differences in judgments and conceptions about critical thinking that arose among the experienced faculty rankers, and the differences between them and the cohort of trained scorers, suggest the complexity of the skills that are understood as “critical thinking.” At the same time, the similarities in judgment led us to believe that we do not need to fix a single definition before creating a means of assessing its change over time. Disciplines will continue to emphasize different aspects of critical thinking, but there is also a good deal of consensus in practical judgments. Even the differences can be useful, when they are thoughtfully articulated. For example, discussions of genre differences in Stage Three of our project have inspired instructors to reconsider writing assignments, differentiating those that invite students to think more critically from those that elicit less thoughtful analyses.

That is the ultimate purpose of program assessment: to improve the quality of instruction and the skills acquired by students enrolled in classes. As we proceed from stage to stage we are learning from our project, and the results reported above suggest that our students are learning too. We believe faculty on any campus could be trained in this method to assess the effectiveness of their educational program, responding to calls for accountability with evidence that their students are learning to think critically.

July, 2007