REVIEWS

A Review of Norbert Elliot, *On a Scale: A Social History of Writing Assessment in America* (New York: Lang, 2005), Richard H. Haswell, ed., *Beyond Outcomes: Assessment and Instruction Within a University Writing Program* (Westport, CT: Ablex, 2001), Brian Huot, *(Re)Articulating Writing Assessment for Teaching and Learning* (Logan, UT: Utah State UP, 2002), Bob Broad, *What We Really Value: Beyond Rubrics in Teaching and Assessing Writing* (Logan, UT: Utah State UP, 2003).

Harry Denny and Pat Belanoff
Stony Brook University

December 2004

The conference building was a soothing fusion of feng shui-inspired angles and curves and a postmodern palette, crisp and calm. Corkscrewing down the stairs to a floor of conference rooms, we arrived at our reading space, an unremarkable room in an otherwise stunning structure. It had the shape of a long rectangle, with windows that looked out onto an atrium void; clusters of tables have been arranged around the room, at which groups of instructors would read three sets of essays—researched arguments, textual analyses, and take-home standardized-prompt essays. Seating was based on assessment traits (critical thinking, genre knowledge, rhetorical knowledge, and mechanic and usage), and each individual was assigned to a table by administrators based on his/her ability and willingness to assess a specific variable. By the middle of the first day and for the remainder of the three-day assessment period – after discussions about pre-selected anchoring essays – groups sat in silence reading and scoring. The monastic atmosphere was punctuated by brief moments of hushed talk; conversation centered on readers and

scoring leaders narrowing in on accurate scores for particular essays. Even these talks, apparently productive and on-task, would get shushed by agitated readers trying to focus on the student writing in front of them--or subconsciously projecting their frustration.

December 2005

Back in our renovated building after three years as a program without a permanent home, small instructor-driven portfolio groups have returned to replace the mass reading events. In a windowless lounge, a group of lecturers and graduate teaching assistants gather. Collections of student papers are tucked into a constellation of colored folders, and they form uneven, askew piles. Talk ebbs and flows as teachers share portfolios, read them and negotiate whether the students they represent ought to pass (or not pass). These conversations are the culmination of a semester-long collaboration—the instructors have met with one another to hash out portfolio contents (the sorts of essays and in-class writing that each would use) and standards, and students and teachers have conferenced and workshopped in various formats, focusing on revision and editing. At the end of the semester, talk focuses on helping instructors' evaluation with the support of their peers; students' textual performance gets married with other contextual cues and factors. Teachers weigh the quality of products with students' learning about process and their effort; debate centers on mitigating factors like language acquisition for non-native English writers or essay prompts that misfired. The meetings, though time-consuming, remain fresh because interaction centers on support, pleading and challenging, problem-posing and searching.

For composition teachers and writing program administrators around the country, variations of these two moments are common experiences. The ritual of instructors responding to and evaluating student writing in isolation is becoming the stuff of academic de rigueur, as institutional, social, and pedagogical change creates pressure for wider assessment of learning

outcomes; consequently, more and more teachers are gathering into rating groups to score student writing. Though fears of corporate managerialism in higher education are well-founded (as evidenced by the havoc wrought by No Child Left Behind federal mandates on K-12 curriculum and pedagogy), institutional needs to document and assure curriculum delivery are reasonable, and writing program administrators and instructors alike are wise to respond pro-actively, lest they become obstacles at their own peril, thus allowing for a vacuum into which testing technologies that lack any local ownership can filter in. Stony Brook University has a long history with writing assessment, and its writing program has been both a cause of celebration for innovation and site of crisis for administration. At each of these moments, the focus for cheers and jeers has been the evaluation of student writing.

As widely published elsewhere, in the mid-1980s Stony Brook moved to portfolios as a means to judge student writing and to foster talk and community among instructors. This turn to group reading also worked to eliminate a longstanding proficiency essay examination. The portfolio included an academic essay (analytic or researched), a personal narrative, and informal in-class writing. As time passed and WPAs changed, the administration of readings grew more complicated, shifting from the small groups and their conversations to a more regimented process involving normed anchor essays and the blinding of student identities. The reliability and validity of scoring the essays and the rhetorical power invested in those technologies consumed ever greater energy and focus, since the program was under pressure to empirically justify its growth and budget through the efficacy of its teaching. Then, state-mandated assessment hit the SUNY system, and local campuses were required to develop their own protocol or use a nationally-normed test (like ETS's Academic Profile or ACT's CAAP). Stony Brook opted for the local approach. Yielding to state pressure to correlate our findings with "national" norms, we moved our portfolio away from talk-invested scoring to an educational measurement-inspired strategy that divested

instructors of context and sought to standardize the assessment of primary traits, reflecting a skills-based approach to writing instruction.

Richard Haswell and his colleagues at Washington State University responded quite differently to state-mandated assessment. Their approach is described from multiple points of view in *Beyond Outcomes: Assessment and Instruction within a University Writing Program*, a collection of narratives, reports and reflections. The WSU system involves a two-essay exam (analytic/analysis and reflective) during placement and at the "rising junior" moment, and it also reviews portfolios from first-year writing and student-selected writing from across the curriculum in the intervening semesters. Students produce twenty pages of finished writing for both portfolios, and faculty from within the writing program evaluate placements and first-year portfolios while instructors from across campus read the writing across the curriculum. WSU uses a two-tier system for each of its ratings, with each execution triaging "obvious" decisions from more complicated ones. For the "rising junior" readings, for example, readers quickly decide between passing and "needing work" decisions, with the latter portfolios plus those that signal "distinction" drawing additional reads. Students who "need work" are placed into additional writing courses and participate in a credit-bearing tutorial operated in conjunction with the university writing center. As part of the book's emphasis on multivocality in the development of assessment regimes, *Beyond Outcomes* includes the critical perspectives of students and faculty alike. Furthermore, the contributors to this volume unabashedly acknowledge poor decisions and problems not yet solved.

Both Stony Brook's and Washington State's more comprehensive systems are part of the long history of the assessment of writing that Norbert Elliot traces with great detail in *On a Scale: A Social History of Writing Assessment in America*. One of the most chilling statements in this book (and one we'd like to dispute but cannot) appears early on: "Testing—with its origin in the European science of psychological measurement—would

become America's unique contribution to education" (4). Elliot moves from the elitism of Harvard's admission tests to the social engineering espoused by the World War I tests of recruits and the establishment of the College Entrance Examination Board on to the devotion to multiple choice tests and then to primary trait and holistic scoring systems and finally to portfolios. Each of these mutations he places in the political and social context of its time. For example, he notes that CEEB's noble intent was to reward merit regardless of social class and ethnic background and move away from the literary foundation of Harvard's entrance tests. We can applaud that while at the same time lamenting the solution: multiple-choice tests of various aspects of language that now we can recognize as assessing almost the same class-based skills as Harvard did.

This book is not easy reading; it often seemed to backtrack as it looks at this history sequentially through different lenses. For example, in the tracing of reliability results in tests across time (pp. 277-92), Elliot retraces historical developments that he has already looked at in terms of the personalities and changes involved in the work of the College Entrance Examination Board in Chapter 4. Nonetheless, the chart in the final chapter of the book is worth the price of the book. With this outline in place, one can see all the interacting threads that Eliot has traced separately in earlier chapters. We suggest that readers may well want to look at this chart first and then read the book.

In concluding, Elliot notes that he has attempted to remain impartial in his presentation of this history. He reflects, "I...wanted to avoid grand systems of interpretative categorization. I did not want to advance an argument, but, rather, wanted to document those historic occasions in which communities met to investigate writing ability" (314-15). Nonetheless, we can read through and behind his account in a number of places (and, to be fair to him, he acknowledges that reporting without interpretation is not possible); consequently, we are not surprised when he concludes, "The more we problematize the construct of literacy, the more the goal of

efficient assessment recedes into the distances; the more we seek efficiency, the more our construct of literacy deteriorates into a postcard's worth of crude common sense" (352). What we end up taking away from this detailed history is that assessment cannot be understood outside the environment in which it occurs. Elliot sets this forth in the context of the history of our country, but we can read this dictum also in terms of our individual campuses: we must assess within the environment where consequences are played out.

Brian Huot's *(Re)Articulating Writing Assessment for Teaching and Learning* sets itself up as an argument from the outset. He asserts that the assessment of writing needs to be rethought as research. At the conclusion, he states this theme quite adamantly: "It is not easy to make any substantive changes in writing assessment practices because we must do more than just change practice, we must be able to disrupt the theoretical and epistemological foundations upon which the assessments are developed and implemented.....we must...learn to look past the technological orientiation of assessment and begin to see it as research" (176).

Each of the chapters in this book takes up this argument from a different position, but none of the argument is polemical. Huot presents in briefer form some of the history covered in the Elliot book to demonstrate the ways in which the assessment community and the composition/rhetoric community have tended to work separately—not even along parallel paths. In harmony with Elliot, he sees validity as the driving force in assessment today, rather than reliability which for so long was the desired goal. Additionally, he sees composition and rhetoric people more concerned with validity as measured by the effects of a test than most psychometricians are. If, Huot notes, writing faculty are to take charge of assessment, they are going to have to learn something about the technical aspects of it, an undertaking that Huot rightfully recognizes many of us resist.

We believe that Huot's book is an essential read for everyone in the field of composition and rhetoric—and not just for those whose scholarship and practice focuses on assessment. All of us in

the discipline need to have under our belts the history he gives us and the rationales he provides for his conclusions. Particularly valuable is his chapter on the importance of reading and response to assessment. For the most part, psychometricians are interested in how evaluators read in order to make them read more like all other evaluators. Huot is concerned that, as readers, we all understand that we're always interpreting and we need to understand how interpretation fits into our assessments. (Broad develops this concept even more fully as we'll note later.) Through arguments such as these, Huot makes a firm link into the literary and cultural-studies approaches but never (and this is crucial) abandons his firm commitment to our equal footing in psychometrics. As he notes, it is the interpretation that is valid or invalid, not the test.

These three books can be linked: Eliot's is a full national history of assessment; Haswell's is a detailing of one program's answer to the call for assessment, a kind of micro-history; and Huot's brings together the history and the implications of this history for campuses. Though he stays on a less specific plane than do Haswell and his colleagues, Huot does present two specific cases for analysis.

Bob Broad's book, on the other hand, begins in a different place as a case study of organic development of assessment technology. Huot makes the argument that concern for reliability is the gate/barrier between education measurement specialists and compositionists. Most of this attention to reliability centers on holistic scoring and the attempt to get teacher/raters to score according to a common rubric. But the process by which they get there forces them to surrender their expertise to an artifical set of criteria, as embodied in an imposed rubric or set of rubrics. In truth, there is no proof that the holistic method does not result in individual variation. Ostensibly, the training process for holistic scoring creates a veneer of like-minded reading, but beneath that surface we do not know how teachers actually read. And this is where Broad's book begins; perhaps a better sub-title for his book would have been *Before Rubrics*.

Broad does his best to crawl into the minds of teachers and help them articulate both for themselves and for him what the grounds of their evaluative decisions are. He records conversations occurring during rating sessions as well as comments made in interviews he holds afterwards with some of the teachers. In truth, some of this becomes rather monotonous, but the results are well worth the effort, for Broad's point is that we need to acknowledge that all of us apply given rubrics in our own way, pushing them here and there, reading them in our own ways, overvaluing some and undervaluing others. Since this is the case, the most logical procedure is to build rubrics from the ground up. First, discover what evaluators give as their reasons for decisions, then categorize them in some way, and create a local rubric. This is exactly what Broad does in his book and exactly what the other three books do not consider: what makes each of us as individuals think a paper is good or poor writing? (This is not a criticism of the other books; they do not undertake such a task within their purposes for writing.)

Certainly this book demonstrates how evanescent and particularized evaluation is. We found ourselves wincing at spots when evaluators pass a student because s/he (in the evaluator's judgment) is ready to move on and fail another because s/he is not ready to move on. One cannot help recognizing how subjective all this is—and yet, deny it if we wish, it **is** subjective. We found ourselves recognizing some of these measures and realizing that we too used them and winced again. Is there some secret wish on the part of all of us to find an absolute measure that would exonerate us in a way from making decisions about students' lives? Would there were some MRI that would simply scan our students and tell us how to deal with them. Huot would answer, and we agree, that assessment is a necessary and valuable part of teaching, and the more we see it that way, the less we will resent assessment and the more useful we will find it. It seems quite obvious that the teachers Broad spoke to were finding assessment talk influencing pedagogical decisions in desirable ways.

Broad then categorizes the evaluative statements he has heard and begins to create a rubric that will be applicable only within the context in which it is created. A number of these categories he sets up on the basis of teacher talk seem quite subjective and interpretative; we can see other ways of grouping them. But, that is not the point: the point is for a program to create the categories it uncovers within its own evaluative talk.

Broad demonstrates for us how best to create rubrics for use in assessment at an individual campus. Haswell and his colleagues describe from a number of viewpoints what a comprehensive assessment program at one school might look like and what goes into setting it up, including the roadblocks and errors along the way. Like Haswell, Broad acknowledges his mistakes—he realizes, for instance, later in his study, something he should have done as part of the earlier data collection, but cannot now recover that ground. But the point here is that both authors in discussing evaluation recognize its complexity: that it's just as difficult to get a program or study exactly right as it is to get a student's score exactly right.

Alas, not many programs are going to put in the time and effort that Haswell and Broad stir up. Haswell has been able to garner the support of faculty throughout the campus who read portfolios on a regular basis; Broad has been able to spend considerable time listening and talking to teachers. Nonetheless, there is something in both books for all of us, something each one of us can pick out and implement even if it isn't the whole pineapple. Nor would these authors even advocate others to create the assessment they have created. We have to do the hard work ourselves.

But these books do offer lessons for our hard work, lessons that we and others are well to heed as we work out assessment strategies. The first and most important of these lessons is that attention to context must drive approaches to evaluating student writing. When assessment comes down the pike, instructors and administrators need to understand, not just battle, the forces promoting it. In New York State, SUNY-mandated assessment grew out of changes in political leadership and conservative

skepticism about instruction and curriculum delivery: was instruction enabling students to learn and develop skills? These forces are not too far afield of movements to control other components of pedagogy in the classroom—the Academic Bill of Rights serves as a powerful example and reminds us of Eliot's contextualized changes in assessment within social and political environments. On local campuses, acting as bulwarks to assessment is not the best response to mandates; even if that were desirable, it is not something we or the authors of these books would advocate. But we do need to impress on those in charge that context is crucial if assessment results are to be focused back into the improvement of the classroom—a goal certainly shared by teachers and administrators of all political colorings.

The second lesson set forth by these authors is that, like the old adage about politics being local, assessment schemes work best when they are organic to the teachers, students, classrooms, and curricula being examined. Rubrics, whether drawn up from collective practice (like Broad) or negotiated by stakeholders (like at WSU), must reflect the values of local instructors, and protocols need to aid in the formative understanding of what marks good and not-so-good writing. As the ones who are most likely to do the heavy lifting (reading portfolios and scoring essays), teachers are at the frontlines of any assessment, and they have the most to gain (and lose) by the results. At Stony Brook, our December 2004 process became disconnected from instructors' actual experiences in classrooms, so the yield, while reliable, lacked validity. Our interpretations and arguments, to borrow from Huot, did not resonate with the values that our teachers possess. Moreover, our students never became a part of the conversation about assessment; the process became so wedded to ensuring sound methods that it became artificial.

The third lesson, stated long ago by Ed White's oft quoted "If we don't do assessment, it will be done to us," is that we must take control of assessment—and, to do that, we must familiarize ourselves with the vocabulary and aims and purposes of

psychometricians. We cannot hold ourselves above such talk if we want to bridge the gap between "us and them."

And, the fourth lesson (and in some ways perhaps the most important) is that assessment can do us a world of good if we design it to feed back into the classroom. After all, one of the purposes of assessment is to make us all better at what we do and that, per se, requires reflection and some objects to reflect upon—in this case, the results of intelligently designed and implemented assessment. We can make it work for us and for our students if we take it on as our responsibility.

Perhaps what's growing out of our look at these books is a substantiation of what most in the field know: that an assessment program is no better than the work and thought put into it and that this work needs to be viewed as a text-in-progress just as student papers are so viewed. We'll never ever get it exactly right—and that means we will always have to be working on it and that's good because it feeds back into our teaching and keeps it vital and up to snuff as times and needs and students change. If we are informed about our assessment history (ala Elliot), aware of possibilities and of input from all aspects of our institutions (ala Haswell), tuned into what our teachers value (ala Broad), and inspired to see what can be changed and different (ala Huot), we can make assessment work on our campus for us, for our institutions, and—most of all—for our students.

These books have helped us at Stony Brook understand better the two scenarios we presented to open this review: each has its strengths and its weaknesses. The December 2004 process adhered to the conventions of psychometrics and educational measurement of writing, and the experience one year later was more faithful to the values of composition studies theory and practice. The earlier approach signaled better science and inadequate regard for local culture, and the recent experiment, albeit lacking empirical data and reproducible insights, possessed deep resonance for our community of writing instructors and students and their learning needs. These books have helped us understand this dichotomy more productively, but have also

provided us a lens through which to being to find our bridge. No one can do that except those of us who teach on this campus. But we must always keep in mind that it is dangerous for us to surrender evaluation to a mirage of true scoring of student writing. At their best, assessment regimes, like any grading protocol, capture student performance at one particular moment. We can learn by assessing that one moment and can turn that learning into numbers and into changes in our curricula and classrooms, but we need to be cautious about believing that that one moment tells us all that we need to know. No assessment, not even the oncoming assessment all of us practice in our classrooms, will ever do that.