Dundee Lackey
Dr. Malea Powell
AL 885: Research Colloquium
October 7, 2004

Assessing Writing

Assessing Writing is described as "...a refereed international journal providing a forum for ideas, research and practice on the assessment of written language" and "...publish[ing] articles, book reviews, conference reports, and academic exchanges concerning writing assessments of all kinds [...] [and] all stages of the writing assessment process, including needs evaluation, assessment creation, implementation, and validation, and test development [...]. They are interested in articles on "...the assessment of writing in the fields of composition, writing across the curriculum, and TESOL (the teaching of English to speakers of other languages), [*Assessing Writing]* welcomes articles about the assessment of writing in professional and academic areas outside these fields ("Assessing Writing" par. 2). Review essays cover "...key issues in the theory and practice of writing assessment" ("Assessing Writing" par. 1).

While the research area here is very focused, "the scope of the journal is wide, and embraces all work in the field at all age levels, in large-scale (international, national and state) as well as classroom, educational and non-educational institutional contexts, writing and programme evaluation, writing and critical literacy, and the role of technology in the assessment of writing" ("Assessing Writing" par. 3). The journal also "... attempt[s] to reflect the concerns of teachers, researchers and writing assessment specialists around the world, whatever their linguistic background" ("Assessing Writing" par. 2), and does consider for publication articles in English about the assessment of writing in languages other than English" ("Assessing Writing" par. 2).

I chose "assessment" as the focus of this week's readings because, although it is an interest of mine, I will not be studying assessment as part of my disseration (I don't think!). My primary reason for selecting this focus is that I feel "assessment, diagnosis, and evaluation" is a necessary field of study for all teachers who grade student writing; it is, therefore, something I hate to leave behind. Some of these pieces focus on basic writers, as, for the first time, I am teaching a "basic writing" course, Preparation for College Writing.

I chose to survey four articles spanning volume 8 (2003) and volume 9 (2004). These articles discuss the idea of assessment in several contexts. James Penny's 2003 article discusses the assessment of

timed high-stakes writing placement exams from an insider's perspective. Yeonsuk Cho's article, which appeared in the same issue, takes us through a study comparing two different types of writing placement exams, one a very traditional "product centered" approach (like that discussed by Penny) and the other a "workshop" approach built on what we know of writing as a cognitive process strengthened by time to think, read, and revise. Icy Lee's article focuses on how ESL teachers mark and provide feedback on errors in student writing. Finally, Condon and Kelly-Riley's article takes us through the use of Washington State University's Critical Thinking Project and the use of the *WSU Guide to Rating Critical Thinking* as a way of designing writing tasks and assessing student writing. These articles seem to accurately represent the breadth of coverage by the journal.

"Reading High Stakes Writing Samples: My Life as a Reader," James Penny

According to the abstract, Penny takes us through life as a "measurement methodologist … training to work as a part-time reader in a high stakes assessment of writing samples from a statewide competency-testing program for ninth graders" (192). He does this, of course, but what you don't get from that statement is the voice he does it in. He simply tells the story, beginning with a caveat entitled "Prologue: Let the Reader Be Warned" which essentially says here there be no quantitative analysis: "you won't find a single numerical measure (though I do conjecture with one correlation), not one equation, no quantitative hypothesis.... This style is quite out of the ordinary for…" Penny, who has published approximately 175 articles as a result of his quantitative research (192).

It's interesting that he chose this very narrative, personal voice for this piece, and particularly in a journal like *Assessing Writing*, whose pages, at least those I've surfed through for this study, seem filled with charts and measurements. This seems appropriate—assessment and evaluation, in general, have always struck me as the math of English, like balancing your checkbook. Nobody really wants to do it, nobody really ever feels completely secure they're not being subjective in some ways—and yet, it must be done. So we try to apply all kinds of numbers and measurements to find ways of doing this. The fact that Penny steps out of his by now, apparently, customary role as a researcher and writer to simply, yet powerfully, tell of his experiences in this area of our field (and right from the beginning to, speaking of money and ads and interviews and all the nuts and bolts of life—he uses this minutiae to provide a slice of life—to humanize what we dehumanize with our numbers.

Somewhat ironically, interviewees must submit to a timed writing sample (doubly ironic considering Penny just told us all about bringing in his Ph.D.--frame and all—to be photocopied. Presumably, he can write. What's with the timed test?) .

> There are many problems facing public education today, and each is worthy of its own dissertation, or more, but I have four pages, twenty minutes, a pencil, and one right hand, so I must choose my one problem, for this one day, for this single essay, while this one soft-leaded pencil points lasts, and until my four fingers and one thumb loose their fine motor control, so alone at this lunchroom table, on this plastic chair, in this deserted dime store mall a half-click north of the interstate leading southward to the promised land, I make my one choice, and that choice, this moment, this interview, this instance, will be the professional status of teachers, or the lack thereof. It was downhill from there to the bottom of the fourth page where I crunched together ten additional lines in the bottom one-inch margin, forming a futile attempt to express an opinion developed through four decades of experience with public education (196).

In this sample, we begin to see Penny's frustrations, and something more of the reasons why he has chosen to make the argument he's building here *this* way, in this voice, rather than documenting (as he so easily could have) a dataset to support that argument. It seems obvious that Penny probably doesn't really NEED numbers for this article, especially for this audience. We all know that "…we should assess students as we train them" (196). The schism between how we actually write and how we are tested as writers magnifies Penny's frustration, even at this "pre-reader-employment" stage, because he had to write the sample longhand, which does not represent his own composing process:

> I composed at the keyboard; I rarely typed from hand written copy, so the psycho-motor skills that permit written composition by pushing and pulling a pencil across paper have long since atrophied. Of course,there's also the business of my penmanship, but that's another matter all together. I depend on a keyboard to write, and frankly that collection of ill-arranged keys has become an extension of my fingers into which I pour my thoughts. In addition, I depend heavily on spelling and somewhat on grammar checkers to fix my mistakes automatically, so I rarely slow down to correct the small errors, and I think that's a good thing, especially when free writing in the early stages of a manuscript. Moreover, I depend on the cut-and-paste facility to make up for my predilection to afterthoughts. Like most folks, I rarely write a paper from beginning to end […]All this thinking left me to wonder about the validity of collecting writing samples in a manner different than many students likely otherwise write" (196).

From here, Penny takes us through the reader training. There is a four-point rubric, with two points passing and two points failing (197). Mechanical features such as grammar, spelling, punctuation and penmanship not to be taken into account (197)

> There is nothing proprietary about the rubric, though I did notice that the NDA precluded me from removing the rubric description from the premises. I suppose the state felt its expansion of the typical and oft-used holistic rubric was unique enough to make it marketable. More likely, my darker side mused, the state officials didn't want the newspapers analyzing the rubric and its use in a public forum (197).

Penny goes on to share with us without violating his NDA by providing a "hypothetical example" of each point level, based on his reading of about 4,000 writing samples over those few weeks (198-9). The samples are hardly necessary; anyone who has been teaching any length of time would, I think, already have encountered this scale. The descriptions, however, help to frame a discussion on how the readers react, and how we define "good" writing.

> On reflection, I wonder how some students could write so well using pencil on paper. Perhaps they are accustomed to the slow process of thought and composure that must precede handwriting, and few of these papers showed little evidence of editing, so I have to assume the students wrote what they composed as they thought of it. I wonder how using a computer might have influenced those papers. My feeling is there would be little improvement in the final rating, though a few papers may have received a higher mark. More likely, the papers would have been longer, surely contained fewer misspellings, and perhaps used better grammar. However, I'm not convinced that the 4-point rubric could often have captured the change a word processor might introduce. More likely, I suspect the students who wrote well on this occasion were just plain good writers... (205)

Penny takes us through special cases, including some that are probably familiar to all teachers of writing, such as the blank page, the 3 page "plea for mercy", the paper that merely recounts the plot, the Spanish paper, the refusal (206-7). He also reveals that there was a procedure in place in case anyone was to receive for scoring a suicide note. He spends a good deal of time discussing assessment failures, both his own and those of colleagues, (213), which leads us into a discussion of reader reliability and, his own personality as a reader ("quick to make decisions") and personality traits that he thinks are helpful (and harmful) when doing this kind of work.

Penny notes that the Myers-Briggs test identifies him as an INFP (intuitive, introverted, feeling and perceiving) (208). He feels that being introverted "...is likely a strong asset for a reader" (208), but notes that "...other aspects of [his] personality should have presented a problem for [him]" (209), since:

> the application of the rubric to a sample did not call for any reading between the lines. It called for no filling in the gaps, no discerning what the writer was really trying to say, and no gleaning hidden meaning. There was certainly to be no reading between the lines, which is where many good writers do their best work. All the reader was to do was read the sample and apply the rubric consistently. Hence, there is no place for the reader to apply personal or professional feelings about the quality of a writing sample. The only thing that counted was the manner in which the sample fit into the rubric defined by the state (209).

He notes that what probably saved him here was the NFP ability to also operate as a STF (sensing, thinking, judging) personality (209-10). Otherwise, he could not have read as rapidly as required, scoring

as many as 100 papers in 90 minutes (which amounts to less than 1 minute per paper, something even he seems to find alarming in such a high-stakes situation) (210-11). He notes here that, although he read faster than average, his accuracy was at 50% (compared to 71% average accuracy) (211). This leads him to question "validity training" in general (211-12).

Penny ends, as a good researcher should, with framing the possibilities for research:

> I'd be remiss were I to neglect the opportunity to suggest research to further dissect the internal workings of the reader in high stakes assessment. Indeed, my brief experience as a reader convinced me of the need to explore the world of the reader to promote a more reliable and a more valid high stakes assessment methodology. Am I correct in my personal observation that personality influences more than just the comfort of the reader? Do other readers, perhaps coming with different personality preferences, develop alternative strategies for the application of a rubric, strategies that might differ from the one I found in my own work? Might those alternative strategies influence the manner in which the reader assigns marks, resulting in differing degrees of rating reliability? Can we dissect and understand the process engaged by readers, and perhaps improve reader training? It only stands to reason that the improvement of reader training would likely improve the assessment (214).

"Assessing Writing: Are we Bound by Only One Method?", Yeonsuk Cho

This article presents a study (undertaken at the University of Illinois at Urbana-Champaign) that "…looked at how the tests from the two different approaches in writing assessment (product-oriented and process-oriented) affected examinees' test performance by comparing both the textual quality of the test essays and the placement results" (165). The author begins by taking us through some background information on the cognitive model of writing as a process and the implications for writing assessment. Cho's argument is one that many have made: that timed tests, especially those which ask students "…to draft a well-organized essay in less than an hour on a topic that the writer may not have though about before" (168) do not provide an accurate representation or diagnosis of students' true writing abilities. Having established the relevance and necessity of the study, Cho moves to the research design, questions, and data.

The student group followed herein was made up of graduate level "matriculated international students" being tested to place them in ESL writing courses (169-70). The author notes that historically this had been done using an EPT timed writing test which had students "listen to a [videotaped, 10-minute] lecture, read an article, and write an essay that follows academic conventions on a named topic within a testing time of 40 minutes" (170). The workshop test was designed to more accurately reflect "real-life

writing" (171). There were "…a number of facilitating activities built into the test. In the workshop, examinees produce a first draft, and, after receiving feedback from other examinees, a final draft. Only final drafts [were] rated for placement purposes. … The running time of the test [was] approximately six hours" (171).

For this study, students would take the "traditional' test as well as a new "workshop based essay test" (171). Because they had to take both tests, they were offered an incentive for participating in the workshop exam study; students "…were told that the workshop results would override those of the EPT if they placed them in a higher course" (173). Analysis of the test results "…indicated that for the majority of examinees, the workshop helped them receive higher placement results, in other words, to be assigned to a higher course" (176). Further, "nine participants who were initially placed into ESL 400 [the "lowest" level of the three] were exempted from the ESL writing course requirement as a result of the workshop. …none of the participants received lower placement results on the workshop than on the EPT" (176).

The new testing format also documented improvement in the "textual quality" of examinees' essays. The study revealed that "…the workshop essays … received high ratings on all the features and aspects except for 'source attribution.' … To explain this result, one of the raters speculated that the longer essays could have provided the raters more opportunity to find examples of improperly attributed information or information that was not cited at all in an essay" (177). Additionally, "…the revised workshop essays in general received higher rating in terms of the content, indicating that the workshop essays had more substantive and elaborated ideas than the EPT essays" (179).

As a "control" measure, the researchers asked content-field faculty to "serve… as an external criterion to evaluate the accuracy…" of the two tests in placing students in the "proper" ESL course (180). The study found that "the EPT placement results agreed with the faculty evaluation in only 41% of the cases … and placed 48% of the students in a lower class than the faculty suggested" (181). By contrast, "the exact agreement between the workshop and the faculty was 48%" (181).

The student examinees were also asked to rate their experiences on the two testing formats. Less than half (47%) of respondents "evaluated the EPT positively (181). In rating "…their *performance* on the EPT, 61% of the students reported that they were not able to write as well as they usually do" (182, emphasis mine). These students were asked a follow-up question "…to specify what elements of the EPT

they believed led them to perform poorly. Of the reasons provided, unfamiliarity of topics (62%) and a time constraint (55%) were chosen most frequently" (182). Students also identified a possible "bias in topics [which] favor[ed] a specific group of examinees over others" (182). In rating the workshop based test, examinees revealed that they "…favored the new writing test," believing it "…a good way to measure writing ability (79%)," and "…provided a more natural writing environment than the EPT (93%)" (182). They also felt "…that the extended time (84%) and the reduced anxiety (79%) helped them perform better on the workshop" (182).

In the final analysis, Cho believes that "…essays produced on the process-oriented workshop test had more elaborated ideas and better organization than their counterparts written on the product-oriented test"; however, "concerning the accuracy of placement results, the results of both tests are discouraging. The ESL writing teacher survey indicated that using the workshop-based essay test, placement of some students still had to be readjusted using results from the proficiency test" (182).

"L2 Writing Teachers' Perspectives, Practices, and Problems Regarding Error Feedback," Icy Lee

Lee's article discusses different types of error marking, from the comprehensive correct-them-all-in-red-ink strategy, to the current notion of "what's best," marking them selectively. Though the literature clearly shows the latter to be a better choice for both students and teachers, Lee's study, which surveyed "…206 secondary English teachers in Hong Kong…" (216), reveals that what we know and what we *do* are not always the same thing.

One difficulty is that teachers, even though we *know* the theories behind selective or minimal marking indicate that "…indirect feedback (ie. indicating errors without correcting them) brings more benefits to students' long-term writing development than direct feedback" (217), the process often short-circuits. In the L2 classroom, part of this, Lee speculates, is because "…students attach a great deal of importance to writing accuracy and are eager to obtain feedback on their errors" (Cohen, Ferris & Roberts, Lee, and Leki cited in I. Lee 217). Additionally, teachers are divided over the *type* of indirect feedback to use. Should we, at times, give direct feedback and model corrections to help learners identify the types of errors they are making consistently? Should we give indirect, coded feedback that identifies, using a code, what kind of errors are there, a process that is often "…cumbersome for the teacher and confusing for the

student" (Ferris qtd. in Lee 217)? Should we mark *all* errors, or just a certain subset of error type? And how do we choose that subset? (218).

For this study, Lee's research questions were:

1) How do teachers give error feedback in the writing classroom?
2) What are teachers' views and beliefs regarding error feedback in the writing classroom?
3) What are teachers' problems and concerns regarding error feedback in the writing classroom? (219).

The study revealed that "...the most common error feedback technique always/often used is 'indicating and correcting errors' (i.e. direct feedback), followed by 'indicated errors, categorizing but not correcting them" (224). Direct feedback was reported by 46% of teachers and indirect coded feedback by 36% of teachers (224).

Teachers justified the choice of comprehensive marking in a variety of ways. The most notable statements here were: "Students prefer comprehensive marking to selective marking," "Teachers are considered lazy if they do not mark all student errors," "It is the teachers' duty to mark all student errors," and "Parents want teachers to mark all errors" (221). These statements reveal that "...though some may thing selective marking is a better idea [,] teachers probably see error correction as their responsibility and feel that it is hard to avoid the job" (221). Despite this--and despite statements like "If teachers don't mark all errors, students do not know what kind of errors they have made" (221)—some teachers *also* felt that "even if [we] mark all the errors, they [students] still make the same types of mistakes next time" (222). Clearly there is a divide between our theorized beliefs, our personal beliefs, and our practices.

Even those teachers who reported using selective marking marked a relatively high percentage of errors (approximately 2/3 and above) (222). There was also some debate over how the error types to be marked were chosen, with "...quite a large number of teachers ... cho[osing] the errors selected on any one occasion on an ad hoc basis" (222) and "a relatively low percentage of teachers ... indicat[ing] that the major principle for error selection was related to students' specific needs" (222-3).

Perhaps the true difficulty in students learning from either type of marking scheme lies in the feedback we give students on their errors *after* we mark them. The majority of teachers reported that "...they would make students correct errors/in/outside class" and go through students' common errors in class (225). 60-68% of teachers reported the first strategy and 72%-78% the latter (225). Very few (from 20-33%, depending on the grade level) conferenced with students about their individual writing and error

types (225). Still fewer (4-9%) made use of error logs (225). This suggests that students are not learning to improve their errors, no matter *what* type of marking we choose, because we are not giving them the tools to learn from their "mistakes." Lee suggests that "…teachers need to be made aware of long-term measures to help students become independent editors" (231).

Overall, "although 91% of teachers think that teachers should provide feedback on errors selectively, in reality, only a minority [24-37%, with rates increasing along with the grade level] are practicing it" (226). It seems that, even when aware that students will learn best by locating and categorizing error types for themselves, "…teachers may be driven by the daily and pressing demands of students, parents, panel chairs, principals, etc. to shoulder the responsibility of error location and correction. The idea of empowering students to locate and correct errors may only reside at the back of teacher's minds" (226).

"Assessing and Teaching What We Value: The Relationship Between College-Level Writing and Critical Thinking Abilities," William Condon and Diane Kelly-Riley

I initially chose this article because it was co-authored by William Condon; he and L. Hamp-Lyons are co-editors of *Assessing Writing*. I found, though, upon reading it, that the article fits nicely into my theme for the week, and that it discusses assessment from another angle altogether. In this piece, the authors examine Washington State University's Writing Assessment Program, focusing on *overtly* designing assignments that use writing to help develop critical thinking, as well as the development of a rubric specifically to measure this and give students *direct* feedback showing them where they need improvement and teachers a way to use this feedback in planning lessons.

The article begins by sharing the history of the General Education program from its reworking in 1987. As part of this retooling, they "…defined a set of objectives for the baccalaureate degree, an integrated set of writing requirements, and an integrated structure spanning students' academic careers" (57). By "…the lat 1990s, … [they] could document improvement in student writing …and student growth within our system and their growth as writers"; however, "our faculty … lamented that students lacked adequate higher order thinking abilities … so we began more systematically exploring the relationship

between writing and critical thinking" (58). They found that "…as an instructor group, they tended to" deem acceptable writing that demonstrated "…accurate information retrieval and summary and did not actively elicit thinking skills in their assignment. These forays led us to suspect that in education praxis there may often be little, if any, relationship between writing and critical thinking" (58), despite our lamentations that the latter is "missing." As a result, WSU set out to develop the *Washington State University Guide to Rating Critical Thinking* in order "…to provide a process for improving and a means for measuring students' higher order thinking skills during the course of their college careers" (58).

The *Guide* "…identifies seven key areas of critical thinking":

- Identification of a problem or issue
- Establishment of a clear perspective on the issue
- Recognition of alternative perspectives
- Location of the issue within an appropriate context
- Identification and evaluation of evidence
- Recognition of fundamental assumptions implicit or stated by the representation of an issue
- Assessment of implications and potential conclusion (59).

The Guide is intended to be used "as a diagnostic measure for student progress, and to provide faculty a means to reflect upon and revise their practices" (64). The idea is for "faculty .,, to take the seven-dimension *Guide* and create evaluation criteria and assignments that suit their instructional styles and disciplinary expectations…" (65).

Along with the *Guide*, WSU developed a set of rating procedures (intended to be adapted specifically for each assignment), which lists "…a six-point scale for each dimension" with faculty choosing "..one of the following levels:"

1) Not evident; can't find it anywhere in the paper
2) Discernable, but not developed
3) Better than 2, but not yet 4. Could be confused, inconsistent, etc.
4) Important to the paper.
5) Better than 4, but not yet 6. May be substantially developed in places, but not throughout the paper.
6) Substantially developed; considered in full complexity; nuanced and sophisticated (60).

Faculty are intended to use this 6-point rubric to develop their own criteria for evaluating student work (65).

As WSU switched over from scoring student work under the Writing Assessment Program (which

consisted of an "…entry-level Writing Placement Exam and [a] junio-level timed writing portion of the Writing Portfolio" to evaluating the same work using the *Guide*, they found an inverse relationship. (61) "In other words, the better the writing, the lower the critical thinking score, but the more problematic the writing, the higher the critical thinking score" (61). This backs ups the earlier finding that "…[their] own writing assessment practice tended to elicit and reward surface features of student performance at the expense of higher order thinking" (61), and "…demonstrates the separate nature of writing and critical thinking" (63). This inverse correlation also:

> …point[s] to what anecdotal evidence has long supported. Oftentimes, raters in our Writing Assessment Program comment that the exams seem to show sound writing abilities, but really contain no critical thought, or are vacuous or superficial. Haswell's research (1991) indicates that when writers take risks with new ways of thinking, often their writing breaks down in structure as the student grapples with a new way of thinking (65-6).

The authors note here a point begun earlier: that just because we ask students to write does not, necessarily, mean that we are asking them to think critically: "writing acts as a *vehicle* for critical thinking, but writing is not itself critical thinking" (66). This leads into an analysis of the (obvious) limitations of timed exams (which itself echoes pieces surveyed earlier by Cho and Penny), and, finally, to a call for studying and improving our practices as a field. The authors recommend first that we examine student work through "two lenses" to develop "…a more complicated portrait of what faculty teach and what students learn" (69). They also point out that we must realize as a field that, although "…writing has an important place in higher education, … it is only one way through which many educational objectives can be achieved;" "Critical thinking is a value that all disciplines want to promote, and it can be promoted through writing, but such promotion needs to be done overtly" (69). Thirdly, "…if we want to compile fuller senses of our students' abilities, then we need to pursue better assessment mechanisms. If we want to be able to use the results of our assessments as a basis for instruction, then we need assessments that yield more, not less, information" (69-70). The authors conclude by saying that this article "…reveal[s] a need to look seriously as educational praxis in higher education, first to be sure that we actually promote the values and competencies we claim to promote, and the second to be sure that any assessment that purports to identify those values and competencies actually does so" (70).

Works Cited

"Assessing Writing." Elsevier.com. 2004. 06 Oct. 2004. <http://www.elsevier.com/wps/find/journaldescription.cws_home/620369/description#description>

Condon, William and Diane Kelly-Riley. "Assessing and Teaching what We Value: The Relationship Between College-Level Writing and Critical Thinking Abilities." *Assessing Writing* 9 (2004): 56-75.

Cho, Yeonsuk. "Assessing Writing: Are We Bound by only One Method?" *Assessing Writing* 8 (2003): 165-91.

Lee, Icy. "L2 Writing Teachers' Perspectives, Practices and Problems Regarding Error Feedback." *Assessing Writing* 8 (2003): 216-37.

Penny, James A. "Reading High Stakes Writing Samples: My Life as a Reader." *Assessing Writing* 8 (2003) 192-215.