7
THEORY IN PRACTICE

In the previous chapter, I presented examples of what meaningful and ethical assessment might look like. These examples, however, have been somewhat disjointed, pieces of practices rather than practices as a whole. Moreover, they have been for the most part hypothetical, albeit drawn from my own and others' experience. In this chapter, I analyze actual practices for the ways in which they reflect the theory I am presenting here, and for the ways in which they extend our understanding of it.

First, I look at published accounts of practices at two institutions: the University of Cincinnati and Washington State University. The former provides a relatively early example of how a university implements a portfolio system in ways that respect the expertise of the program participants. I start with this one both because it is one of the earliest examples and because the description of the process is sufficiently detailed to allow for a fairly detailed analysis. Then I turn to Washington State's program, which I have already analyzed using different lenses each time in chapters three and four. In both cases, I argue the theory is the problem, but I do not believe the practice is. There are other ways to understand the same practice that, I would argue, are more productive. In this chapter, I look at the Washington State example through yet another lens; here I analyze it in terms of meaningfulness and ethics.

Both analyses point to strong ethical practices, as I am using that term, but my analysis also shows a lack of attention to the criterion of primary substance, part of meaningfulness, in the literature about these programs. To extend my analysis of the use of primary substance, I then examine two studies that look at what teachers value when they assess student work. The first is Bob Broad's (2003) development of Dynamic Criteria Mapping to examine the particular values of the faculty at "City University." The second, a pilot study I ran at North Carolina State University, uses decision logics to analyze the conversations of teachers during mid- and end-of-term evaluations. In both cases, the studies find that what teachers talk about when they are evaluating writing is not necessarily what appears in program standards.

It is interesting to note that all of these models use portfolios in some fashion. I have chosen these descriptions not for their use of portfolios—

although portfolios, I would argue, lend themselves more readily to theoretically sound types of meaningfulness than do many other assessment instruments. Instead, I have chosen them for their attention to the process of developing a specific assessment praxis that (however unintentionally) illustrates the principles I am developing here.

PORTFOLIOS AT THE UNIVERSITY OF CINCINNATI

The account of the process of establishing portfolio assessment at the University of Cincinnati—published by Marjorie Roemer, Lucille M. Schultz, and Russel K. Durst in two essays—provides an early example that demonstrates in a more comprehensive way some of the key elements of the principles of assessment I am outlining here. The first essay, "Portfolios and the Process of Change" (Roemer, Schultz, and Durst 1991), briefly describes the decision-making process that led the authors to consider program-wide portfolio assessment and then discusses in more detail three pilot studies using portfolios with different groups of instructors. The second essay, "Portfolio Negotiations: Acts in Speech" (Durst, Roemer, and Schultz 1994), describes the on-going process of negotiation that portfolio assessment requires.

In 1987, Roemer, Schultz, and Durst introduced the idea of portfolios and by 1989, after a series of pilot studies, had decided to adopt them program-wide to replace a single-sitting impromptu exit exam. This decision to change procedures was a committee effort, including "both writing and literature specialists, full- and part-time faculty, and a graduate assistant," a respectable cross-section of primary evaluation participants. The desire for change was based on a combined dissatisfaction with the existing exit examination and interest in the value of portfolios as described by literature in the field. Among the primary purposes they articulate for exploring and finally establishing a portfolio assessment, the authors focus on pedagogical value. Portfolio assessment at the end of the first quarter of their three-quarter sequence provides "remediation" for those students who need it at the beginning rather than at the end of the sequence (Roemer, Schultz, and Durst 1991, 456–57). In addition, the mid-semester dry run they implemented provides students with feedback on their overall performance and ability (Durst, Roemer, and Schultz 1994, 287). Instead of serving the purpose of determining proficiency or competence, then, the portfolios as designed provide interim information for both students and teachers that allows for additional instruction, practice, and effort as needed. The pedagogical purpose of this assessment is specific to the context; there is nothing wrong with proficiency or competency examinations per se. Shifting to portfolios alone would not have fulfilled the authors' pedagogical purpose. Moving the assessment from the end of the sequence to the end of the first quarter made the pedagogical goals of the evaluation primary.

The authors are also explicit about their secondary purposes. While they discuss several, the most significant of these for my purposes is the value of dialogue or "speech" for "teacher training and professional development" (Roemer, Schultz, and Durst 1991, 467). Their portfolio assessment procedure is structured around trios: discussion groups of three teachers—with larger meetings in groups of twenty for "norming" purposes (Durst, Roemer, and Schultz 1994, 288). In the discussions, teachers tend to focus on criteria; the authors see this faculty development as a significant advantage, considering that their teaching staff consists of teaching assistants and adjuncts as well as full-time faculty. New teachers gain experience in evaluation while experienced teachers can spend more time discussing complex, recurring issues such as "dialect interference" and overly writer-based prose (Roemer, Schultz, and Durst 1991, 467).

While the authors began the process of changing assessments with the primary purpose explicitly in mind, the secondary purpose of teacher training emerged during the course of the pilot studies. They developed the pilot studies in an attempt to ameliorate the effects of a "top-down imposition" which discourages rather than encourages teacher investment (Roemer, Schultz, and Durst 1991, 463). Between the pilots and the continual solicitation of teacher input, dialogue emerged as a central theme, so much so that by the second essay, the value of negotiations in speech surrounding portfolios becomes the central point of the essay. Dialogue, the authors suggest, not only promotes discussion about standards, but also ownership of the program (457).

Although the authors spend significantly more space on the "purposes" of their program, they do briefly address what I have called "substance." They argue that portfolios better reflect the philosophy and curriculum of their writing program, "fitting [their] emphasis on process, multiple drafting, the development of self-reflective powers and encouraging students to take more responsibility for their own growth as writers" (Durst, Roemer, and Schultz 1994, 287). They argue that the procedure "promoted high standards" (Roemer, Schultz, and Durst 1991, 467), although they do not explain what those standards are. Overall, while a few substantial specifics are offered at various points throughout the essays, the authors do not focus on "content" issues, so the primary substance of the assessment remains vague. They do, however, tie the process of negotiation, a secondary substance, to their secondary purpose of teacher training. That is, negotiation is integrated into their assessment procedure as a substantive way to arrive at their secondary goal of faculty development and teacher training.

While the discussion surrounding what I would call issues of meaningfulness is minimal, the procedure described by Roemer, Schultz and Durst exemplifies ethical assessment as I am using that term. In both essays, the authors mention that their primary reason for adopting portfolios was to focus on students' writing development over time, which they claim is the appropriate emphasis for assessment in their program (Roemer, Schultz, and Durst 1991, 456; Durst, Roemer, and Schultz 1994, 287). In "Process" they expand on this idea: "Students would be judged on their best work. The extent of a student's strengths and deficiencies would be more fully documented and explorable in the portfolio, and judgments would be more defensible" (Roemer, Schultz, and Durst 1991, 456). Although "consistency"—the hallmark of reliability—is offered as a consequence of the procedure in the earlier essay (467), the characteristics of the portfolio assessment described here point to an intention to give each portfolio full consideration. Consistency, then, would be less a result of training than of thorough discussion.

The documentation, exploration, and defensibility of the assessment decisions that the authors describe are not simply an automatic consequence of portfolios, and indeed they point out that portfolios could easily become as rigid as impromptus had at their institution (Roemer, Schultz, and Durst 1991, 457). Instead, it is group negotiation of assessment decisions that offers depth and complexity of information about student writing. The authors conclude that assessment is reading, and that group negotiations "can lead to new interpretations, changed positions, or . . . 'attitude entrenchment'" (Durst, Roemer, and Schultz 1994, 297). The goal of these negotiations is not consensus in the sense that "norming" suggests, where all must come to agreement. Negotiations here allow for differences, and while the authors wonder about the resulting "indeterminacy of the freshman English grade," they suggest that judgments about writing—like anything else in a postmodern world—do not obtain the status of Truth (Durst, Roemer, and Schultz 1994, 298). What is "ethical" about the procedure described here is that polyvocality is integral to the process and to developing a full consideration of student writing.

The procedure outlined by Roemer, Schultz, and Durst is not a full illustration of the theoretical principles I am advocating. I have already noted the limited attention to substance. Furthermore, although there is some mention about the possibility of including students in the evaluation process, it appears without much discussion as part of a transcript; ultimately the procedures were negotiated between the administrators of the writing program and the teachers only. This may be a step in the right direction for developing a limited speech community, but the absence of student voices renders it rather too limited. There is also no discussion of other stakeholders: deans, employers, teachers of more advanced writing courses, and the like.

Still, this program illustrates many of the most important principles of meaningful and ethical assessment that I have outlined above. The procedure develops through community discussion and integrates dialogue into the process of assessment. The purposes and the procedure are internally consistent with one another, and the substance of the assessment appears to reflect a social constructionist paradigm (although discussion is too limited to be clear). This procedure, more than fourth generation evaluation and more than the expert reader model, moves beyond the principles of educational measurement theory toward an assessment drawn from the values at work in composition studies.

ASSESSMENT AT WASHINGTON STATE UNIVERSITY

The Washington State University's system consists of a sequence of assessments and instruction designed to guide students through writing at the university from their general education coursework through their major. The process begins with an impromptu placement exam for first year composition, followed by writing instruction in introductory composition and in the general education curriculum, followed in turn by a mid-career portfolio assessment to determine if students are prepared for Writing in the Major courses. Subsequently, students take at least two Writing in the Major courses, with tutorial support and additional coursework as needed, before they submit to a final assessment within their major department(s).

For the purposes of my analysis in this chapter, the two-tiered assessment used for placement and for the mid-career portfolio is the most germane part of this system. I describe this process in chapter four, but I will briefly summarize it here. In the two-tiered process, the object of assessment is read first very quickly to see if the assessment indicates an "obvious" placement in either English 101 (in the case of placement) or in the "Pass" category (in the case of the portfolio). As Haswell puts it (talking specifically about placement), "[t]he emphasis was on *obviously.*

If the reader found any reason to question placement of the student into [regular freshman composition], wavered in any way, then the sample passed on to the second tier for a different kind of reading" (2001b, 42). While the first reading is quick and focused on a single decision, second tier readings are "deliberate," with an emphasis on thoroughness (2001b, 43). As many readers as necessary read each object, they use all available information, including knowledge of the students' backgrounds, and the readers consult with each other about their decisions as they feel necessary. Instead of the blind, trained readings of holistic scoring sessions, both tiers in this procedure rely on the expertise the readers already have to arrive at an appropriate judgment. The primary purpose of the placement exam is assigning students to the best course for their instructional needs, arguably the primary purpose of any placement exam. And the primary purpose of the mid-career portfolio is to determine if students need additional support in their writing as they do the upper division writing in their majors. These purposes are obviously first and foremost.

The assessments in this program as a whole are designed specifically to identify what kinds of instruction and assistance students need in their writing throughout their college careers, not only at the freshman level. Thus, the program as it stands is inherently cross-disciplinary. The mid-career portfolio, for example, requires that students do a great deal of writing throughout their general education coursework in their first two years. This writing cannot and does not occur only in English courses, nor is it pushed off until students take upper division WAC courses. Consequently, students at WSU produce a significant amount of writing; Condon estimates between 100 and 300 pages of writing per student during their careers (2001, xv).

Primary purposes are those that affect students most directly. One of the explicit purposes of the WSU assessment, however, crosses between what I have called primary and secondary purposes. As Condon describes it, the assessments serve as an integral part of the curriculum, which "features an alternation between assessment and instruction, and the assessment is designed to flow from the instruction and, in turn, to support it" (2001, xv). That is, each assessment is specifically connected to instructional plans and goals for the student. Students do not "fail" the mid-career portfolio, for example. Instead, they are designated as needing work on their writing and required to take a support course (Gen. Ed 302) in addition to the normal two Writing in the Major courses (2001, xvi). And the instruction in the general education curriculum is designed so that students produce ample writing to demonstrate their current educational status and needs as of the mid-career portfolio. This purpose bridges the distinction I have made between primary and secondary purpose because, while it does serve curricular goals primarily, it has a significant and clear effect on students and their progress through the program.

There are a number of other secondary purposes that Condon, Haswell, Wyche, and others identify throughout the book. Condon points out how the program brings faculty together from all over the university (2001, xv). Haswell and Wyche talk about the need to use teachers' expertise "rather than just their obedience to a holistic rubric" (2001, 15). Richard Law, a director of general education at WSU and a member of the original oversight committee for this program, discusses the ways in which the program encouraged and supported reforms in undergraduate education (2001, 11). These secondary purposes focus on the effects that the assessment has on faculty and the program of education.

The authors of *Beyond Outcomes* spend more time than most talking about the primary substance of their assessment. Specifically, Haswell and Wyche emphasize the need for the exam to be linked to the content of the course, which they describe as follows:

> In the last three years, [English 101] had been revamped to focus on academic writing, with emphasis on critical thinking, college-level reading and library skills, and computer literacy. It had also acquired a multicultural slant with its integration into the new general-education requirements. English 101 was conceived as a mainstream course for the great majority of freshmen. There they would face instructors, almost exclusively graduate teaching assistants, all using the same multicultural reader and giving three common assignments: responses to readings, critiques of cultural events on campus, and essays utilizing secondary sources. The placement exam would have to allow us to judge whether students were ready for such a course. (2001, 17)

In identifying the course content and objectives, and specifically tying those to the assessment, Haswell and Wyche define the primary substance of the exam: academic reading and writing, use of source material, computer literacy, and so on. These objectives echo the outcomes developed by members of the Council of Writing Program Administrators (1999), adapted for the particular student body at WSU. And the prompts and procedures used in the placement exam seem to do much of the work that they describe here, clearly linking the course and exam (with computer literacy an exception). Because the authors of the program wanted to look for complex abilities, they developed a complex prompt which relies on interchangeable readings and rhetorical frames that each ask for a somewhat different and sophisticated approach to the reading (Haswell 2001c, 212–13). Moreover, the exam encourages at least some level of planning because students are given the prompt first and the bluebooks 15 minutes later (Haswell and Wyche 2001, 20).

Haswell and Wyche define the course and describe the content of exam that feeds into that course, but it is not a clear match. Aside from the absence of computer literacy (an absence more likely due to logistical constraints more than anything else), the exam does not clearly test academic reading and writing and use of source material; the students read only short passages, and there are no directives to cite anything, for example. What it does do, according to Haswell and Wyche, is provide a "diagnosis of future writing potential. While the holistic forces a comparison of an actual student piece with an ideal and, therefore, highlights the negative, our diagnostic reading would force a prediction of a student writer's success given several different future paths of instruction" (2001, 19). In other words, the primary substance that WSU's placement exam is not the existing ability to do academic reading and writing, for example, but rather the *potential* for students to do that work. The question that the assessment addresses is not "what abilities is this student lacking?" but rather "what is this student ready to take on?"

The dominant paradigm in composition studies, social construction, is not as clearly present in the substance of this exam or the assessment program. While there clearly is collaboration in the assessment procedure, there is no specific mention of principles or practices such as the social construction of knowledge (is the use of source material taught and assessed in terms of how the student is making knowledge with the sources?), or collaborative writing (are students encouraged to do/submit collaborative projects in any of their portfolios?). The absence, however, is hardly damning and is more likely a result of the focus of the text on the development of the program and its processes, rather than on the content of any given examination. And, as I will discuss in a moment, we do see these principles at work when we look at the principle of ethics.

If secondary substance can be thought of as the content that supports secondary purposes, the programmatic support for faculty seems key to this principle. By relying, for example, on the expertise of the teachers in the placement exam rather than on norming, the assessment supports the pedagogical values those faculty members hold. This kind of connection helps ensure that the test will not subsume the course, since those who are teaching the course are those who make the judgments in the assessment. Moreover, the inclusion of faculty from disciplines throughout the curriculum, particularly in the mid-career portfolio assessment, provides campus-wide faculty development through on-going conversations about writing.

WSU's program of assessment and instruction illustrates the idea of a limited speech community, as I have outlined the concept, in more detail than the Cincinnati program does. The initial oversight body responsible for developing the assessment was the "All University Writing Committee," which included faculty members from all over the campus, and which was chaired by a former composition director (Haswell and Wyche 2001, 14). However, Haswell and Wyche argue that this committee "seemed distant from those who would ultimately maintain the [placement] exam and be most affected by it" (2001, 22–23). In response, they brought together "the director of the Writing Center, the administrators of the undergraduate composition program, the leader of the Writing Across the Curriculum (WAC) program, and several writing teachers" to explore the courses, the students, and the possibilities for placement (2001, 23). In so doing, the program leaders established the kind of distinction between full and limited participants that I describe in the previous chapter. That is, they put those most responsible—those who would be held accountable—at the center of the planning and development, relegating those who were interested stakeholders to a supporting role.

Beyond Outcomes also includes a description of the roles of these limited participants in the fourth section, which focuses on the mid-career portfolio assessment. Covering students, faculty, and administrators, the chapters in the section demonstrate how these various groups participate in the maintenance of the program, usually through providing feedback, participating in the assessments and spurring reflection.[56] These stakeholders do not have the power or authority, however, that the full participants do. Their feedback helps initiate and support decisions about changes to the program; for example, student feedback led to the requirement that Gen. Ed 302 be taken concurrently with a Writing in the Major course (Nelson and Kelly-Riley 2001, 157). But those at the center, the program administrators and teachers, made and continue to make the key decisions and to perform the assessments themselves. Central to the way that community works in this assessment is the role of and reliance on expertise. The use of teacher expertise reinforces the value not only of the assessment decisions themselves, but also of the teachers' knowledge. Moreover, the program includes teaching assistants in this group, allowing TAs to make low-stakes decisions, such as the first tier placement decisions. In both these features, the program bases full participation first on expertise, rather than on status within the university.

Earlier I argued that disciplinary knowledge and principles of social construction do not appear in a significant way in the substance of the assessment. They do, however, appear in the process. Tier-one readings are quick and involve make a single decision, based on the readers' expertise: does the student belong in English 101? The tier-two readings, however, are "deliberate"; decisions are made by "as many [readers] as it takes" as part of a "group consultation" (Haswell 2001b, 43). The process for both readings is social, and the decisions are reached as a part of a community. At the second tier, this is obvious: teachers talk freely about the decisions they are making. But the use of expertise in both tiers indicates a reliance on communal knowledge; expertise is born of knowledge of the discipline and experience teaching in the program.

The development of the program similarly reflects social constructionist principles. Haswell and Wyche talk about the influence of assessment literature in the design of the program and its instruments and procedures (2001, 18–21). Locally, limited participants from a range of disciplines participated in the program's development and are involved in its continued support. And ongoing program review depends on those outside the program as well as inside. More importantly, the authors in *Beyond Outcomes* recognize the significance of context to their project. They claim that assessment must be local (e.g., Condon 2001, xiv; Haswell and Wyche 2001, 14), and Haswell and Wyche identify one of the guidelines for any assessment program as "let the local scene shape the examination, not the other way around" (2001, 16), echoing principles raised by Huot and the CCCC Committee. For example, they describe the influence of their own student population on the exams: primarily traditional students at a residential university in a rural location with few students with "severe writing problems" (16). This would not describe the institution where I currently work, nor where I finished my doctorate. Nor do they assume the context is static. Haswell explains that "[j]ust as the test arose out of mutable local conditions and was shaped by them, it should continue to be shaped as those conditions change" (2001b, 40).

This acceptance of the contingency of the project perhaps marks this program most clearly as reliant on social constructionist principles. Nowhere in the text does the reader get the sense that the authors are offering *the* solution to the problem of assessment. Bill Condon calls it an example of "third-wave" and "fourth-generation" assessment, referring to Yancey's and Guba and Lincoln's work respectively (2001, xvii). This claim seems warranted to me, but I would go further, to claim that this program is a strong example of meaningful and ethical assessment.

Using this lens, I believe, we can better understand the workings of the assessment program as a whole and its relation to contemporary thought in composition studies than we can through the lens of psychometrics or even through Haswell's own justification via categorization theory.

LOOKING FOR SUBSTANCE

Both the previous examples provide strong examples of ethical assessments, but both also point out the need for research into the primary substance of writing assessment. On the one hand, it seems odd that our discipline is lacking in this area; after all, it would seem logical that one of our primary concerns would be the content of our courses. On the other hand, we have this knowledge, in part. It appears in our course objectives, our scoring guides, our standards, and our rubrics. It just is not often part of our scholarly record.

Bob Broad makes this point in *What We Really Value: Beyond Rubrics in Teaching and Assessing Writing* (2003), where he describes his research into what I am calling the primary substance of writing assessment. Broad argues that the values typically defined by standards and rubrics—usually some variation on "ideas," "form," "flavor," "mechanics," and "wording" (2003, 6)—do not match what teachers actually value as they actually assess student writing. To get at these actual values, he proposes a process he calls Dynamic Criteria Mapping (DCM), in which members of a writing program can record and analyze their own values as they appear in communal evaluation sessions, such as "norming" or, more appropriately, "articulation" sessions. "Articulation" is the name that Broad suggests both here and in "Pulling Your Hair Out" (2000), because, he argues, "norming" and its synonyms focus only developing agreement among graders, while articulation allows for both agreement and disagreement (2003, 129; 2000, 252). Evaluators, he rightly points out, do not always agree, and their disagreements are often as useful in understanding an assessment as their agreements.

Through his analysis of assessment at "City University," Broad identifies 47 textual criteria, consisting of 32 textual qualities ("aspects of the reading experience") and 15 textual features ("elements of text"), 21 contextual criteria (aspects of the assessment situation and the students' work in class), and 21 other factors, including relations, comparison of texts, the evaluation process, evaluating teaching, and so on. He is able to flesh out concepts that appear in rubrics, such as "significance," through synonyms used by the evaluators, in this case "complexity," "heart," "goes somewhere with it," and "depth" (2003, 40). His process also brings to the fore concepts that influence our evaluations, but do not usually appear in our rubrics, such as a student's progress over the course of a semester. The result is a much richer understanding of the values than "the standard, traditional, five-point rubric, by some version of which nearly every large-scale assessment of writing since 1961 has been strictly guided" (2003, 6).

DCM also helps Broad identify elements on which the program needs to work. In the case of City University, for example, Broad finds that teachers in the program do not have a clear understanding of the role that fulfilling the assignment should play in the assessment (2003, 133). He also finds that mechanics dominated the discussion, in spite of the fact that the program's documents indicate that it is merely one value among many others (2003, 62). Rather than obscure or smooth these disagreements over as a rubric-based assessment would, Broad's method highlights the disagreements so that the program can decide how to manage them. Leaving them open as disagreements is implicitly one of the options.

Broad argues that the purpose of rubrics is to simplify and standardize assessment, that they serve the purposes of efficiency and convenience in what is and what should be a complex and messy process. Broad chose City University in part because the program did not use a rubric. Instead, the program relied on a combination of mid-term and end-term norming and "trio" sessions. In the norming sessions, members of the program debated reasons for passing or failing student work, while in the trio sessions, groups of three instructors evaluated all C or below work for their actual students, what Broad calls "live" student work. Because the program did not use a rubric and instead relied specifically on conversation, Broad felt he was able to get a better representation of the teachers' actual values. Those values represent the primary substance of writing assessment.

In a pilot study at North Carolina State University, Dr. Jane Macoubrie and I found results similar to Broad's.[57] The pilot was designed to test the waters for portfolios in the first-year writing program at NC State and to explore the possibilities of a dialogic assessment procedure based on the theory I am espousing here. Two groups of three teachers participated in the project, assigning portfolios in their classes and commenting on individual papers without grading them. To evaluate these portfolios, the groups exchanged materials, read them in advance, and met to try to reach agreement about a grade for each of them. The teachers were told explicitly that there was no set procedure for performing the evaluation, and that they would set their own plan and pace—although, of course, they had to complete the work in time for final grade submission. They were given instructions to try to reach consensus on the grade range (e.g., A-range, B-range) if not the actual grade of each portfolio, but because of the exploratory nature of the pilot, they were also told that if they could not reach agreement, the classroom teacher would have the final say.

The group sessions were videotaped, and the data analyzed using Macoubrie's concept of decision logics. Much of Macoubrie's research has been on the decision-making process of juries—a situation analogous to dialogic grading, in that juries are required to reach consensus in a finite amount of time, and that they talk about their reasoning and their decisions as a normal part of the process.[58] Specifically, Macoubrie's process analyzes dialogue for two things: (1) patterns in the discussion of substantive issues—called "decisional topics"—and (2) patterns in the justifications used to support particular decisions regarding those substantive issues. In a decision logics analysis, taped dialogue is transcribed and the transcriptions are coded according to local topics and justifications. The local topics are then grouped into more global, decisional topics. For example, in the case of our pilot, the global topic "development" consisted of local topics such as use of examples, definitions, support, evidence, illustrations, anecdotes, contextualization, and so on. The complete list of decisional and local topics appears in Table 1. I directed the pilot and served as one of the teacher-participants; Macoubrie coded the data and tabulated the results, a process I did not participate in.[59] We both worked on interpreting the results.

Like Broad, we found decisional topics that did not appear on the list of program standards. While our list of topics is certainly shorter than Broad's, ours includes topics different from those raised in the City University study: e.g., the quality of the student's research and the student's ability to understand their source material, which is a significant part of the curriculum in the course. Some of these topics are certainly familiar territory and much of the list draws from the existing program standards: focus, development, organization, style, and grammar and mechanics (Freshmen Writing Program 2002). Even though teachers were given no instructions to do so, most—myself included—referred to the standards during the grading process. One of the two groups used them explicitly throughout in their decision making, treating them expressly as a rubric. But even with reference to the standards, teachers still used criteria not explicitly part of the program to make their decisions: "I gave it an A because he was well focused on causes and effects. A brief proposal at the

TABLE 1

*Decisional Topics, All Teacher Groups, with Local Topic Information**

Decisional Topics	*Number of Comments*	*Percent of Comments*	*Local Topics Used*
Grammar & Mechanics	277	16	verb tenses, pronoun use, spelling, sentence structure, punctuation, syntax, paragraphing, MLA, proofreading, capitalization
Focus & Organization	254	14	focus, large-scale organization, format, introduction, conclusion, criteria not separate, flow
Development	234	13	examples, definitions, support, evidence, illustrations, specifics, anecdotes, good discussion, contextualizing, scenarios, conclusions not supported
Argument Quality	196	11	argument, reasoning, analysis, insight, logic, thinking, analysis vs. opinion, prominent claims, argument not clear, unclear point, rebuttal, relationship between ideas, inconsistent facts or claims, wacky interpretations, generalizations
Style	172	9	style, word choice, voice, wordiness, accuracy, repetition, clarity of language, emotion, moving, eloquence, exciting, humor, information not clear, awkward phrasing, parallelism
Research Quality	140	8	references, sufficiency, quality of sources used, bibliography, documentation
Citations or Quotes	127	7	handling of citations, use of quotes appropriately
Local Coherence	111	6	transitions, paragraph coherence, choppiness, connectivity, rambling, cohesiveness
Intellectual Challenge	86	5	challenging topic, intellectual interest, subject attempted, great/poor topic, avoiding thinking
Improvement, Revision, Self Evaluation	82	5	improvement, revisions, addressed problems, self evaluated honestly/critically, self judgment
Content Understanding	51	3	obvious misunderstanding of content analyzed, comprehension of ideas/concepts
Assignment	33	2	part missing, did what asked to do, short, application of prompts, criteria, review, summary/response, no memo
Audience Awareness	16	1	awareness of an audience and their needs
Plagiarism	9	0.5	word choice or phrasing suggests lifted sentences or worse
Total	*1788*	*100.5%*†	

*Table developed from tables originally developed by Dr. Jane Macoubrie, North Carolina State University, Dept. of Communication.

†Does not add to 100% due to rounding.

end. He'd done much revision of his introduction and documentation was much better and his conclusion was clear." This kind of mingling of standards-based topics—focus and organization—with topics not on the list—improvement—was common. And this level of explanation was also common, a point I will return to in a moment.

This "deviation" from standards intensified when it came to reaching consensus on difficult cases. For most of the decisions, discussion was brief and grades for each of the group members were within a letter grade of each other. Reconciliation was easy in these cases. But for about 17% of the decisions (26 out of 152), the grades were more than one letter apart. Ultimately, both groups were able to reach consensus on all portfolios, but when they could not immediately do so, they did not simply return to the original decisional topics. Instead, they added discussion about their own differences in approaches to assignments (as opposed to the student's understanding of the assignment as in the initial decisional topics), the student's improvement over the term, and the use of grades for motivational purposes, among other factors. The complete results appear in Table 2.

TABLE 2

*Reconciliation Means Across Grade Levels, By Case**

Reconciling Decisional Topics	*Number of Cases*	*Percentage of Cases*
Assignment Clarification	6	23
Improvement, Self Evaluation	6	23
Grammar, Style, Technical	6	23
Argument, Thought	3	11
Motivational Grades	3	11
Attendance	1	4
Repeated Plagiarism	1	4
Total	*26*	*99%*†

*Table developed by Dr. Jane Macoubrie, North Carolina State University, Dept. of Communication.

†Does not add to 100% due to rounding.

Beyond the original decisional topics (Improvement, Self-Evaluation; Grammar, Style, Technical; and Argument, Thought), teachers added what Broad would call "contextual criteria." For example, in one reconciliation, when the two non-teachers graded the portfolio harsher than she had, the teacher revised her grade down, acknowledging that "this student is one of the few students in this class who was there all of the time, and actually read the material in advance, and actually participated in the class as opposed to most of the students. I know that I'm pushing it here. I don't have any trouble dropping it down to a C+." In another instance, the non-teachers successfully argued that the student was not ready for the next course in the sequence, and the teacher agreed to a lower grade than he had initially given the portfolio. Non-teachers also served as a counter-force in the other direction. When one teacher had a problem because of a student's behavior in class and was grading inconsistent with the other two evaluators, she raised the grade based on the textual reasons provided: "some good audience awareness . . . style's okay, but not good." In several cases, instructors, sometimes at the suggestion of their group members, raised or lowered a grade, particularly at the midterm evaluations specifically to motivate a student.

More often these difficult cases resulted in prolonged discussion about broader issues. For example, one instructor used a personal experience assignment that the others did not, and both other instructors commented on their difficulty in evaluating the portfolios that included these assignments:

> "I gotta say, one of the problems I have with the entire set of portfolios from yours is I can't do personal experience. I don't assign it and I don't like reading it. I feel very much like a voyeur in students' lives. . . . On my notes here it's like I'm back and forth on grades. Because I really, I just, how do you tell somebody who's . . . "
>
> "Pouring out their guts . . . "
>
> "Yeah, that this is a D. Your guts are only worth a D."

Evaluators had similar conversations about summary/response assignments and papers where the student's thinking was strong, but the language was weak and the paper did not meet the traditional expectations of the course as a result of a teacher's assignment.

More interesting, however, is the fact that regardless of the criteria used, the evaluators did not often give each other concrete reasons for their decisions. They said things like "The summaries are detailed, sometimes too detailed. Does demonstrate knowledge of the text. There are grammatical problems and responses underdeveloped." And everyone nodded and went on to the next person's list of topics to support their grade. Sometimes, evaluators said little more than the grade they had tentatively assigned. But, upon occasion, the comments would become more detailed:

> "And I said the first five sentences of paragraph two were awkward."

> "I got paragraph two, yeah."
>
> "I said some sentences should be separated into two, and others need smoother transitions between the text and the articles. But I thought it was well researched."
>
> "I thought the content was very strong. And the only thing I have in this category, and I was mostly on the first one, so . . ."
>
> "The only comment I have on number three is I agree, it's hard to keep from bouncing back and forth and I felt that she did do some of that."

In the first two comments here, the evaluators point to a specific problem in a specific location, but this kind of specificity is rare in the data, and even when they do get this specific, the groups quickly drop back into general statements, in this case about research and content. When the evaluators return to general comments, their justifications become obscure. In this case, one reading of the transcript could indicate that the two evaluators were talking about two different criteria: research and content. A different reading, however, could argue that research and content refer to the same thing. The comments following this point in the transcript do not help determine the meaning; instead, they take up the topic of organization. Even where the evaluators were using the same words, they did not seem to be communicating about their values, particularly when they listed criteria. I was there, and I am not sure that we all understood each other, even though we appeared to.

This problem with justification and reasoning becomes glaringly obvious when someone from outside the discipline listens in on our conversations and reflects them back. Because my colleague was from Communication, she and I share some of the same vocabulary. She requires writing of her students and talks to them about the importance of audience and purpose; she expects their grammar to be reasonably correct. But when she listened to and watched the taped sessions, she pointed out that while everyone was nodding, even about something as concrete as grammar, they may not have meant the same thing. She could not tell. And the problem was worse in more complex categories. One teacher's "organization problems" could have been issues with the structure of paragraphs, while another's could have been a mismatch between the thesis and the rest of the paper. But the evaluators rarely provided the kind of detail that would help her decide. Nor, in my review of the transcripts, could I.

As a result, there was not enough data for Macoubrie to develop the second part of her decision logics analysis: patterns in the justifications used to support particular decisions regarding decisional topics. All she had from us were the decisional topics, but not the justifications, and without the justifications, she could not thoroughly analyze our reasoning during the decision-making process. In part, the logistics of the pilot got in the way. What we recorded were the midterm and final grading sessions, when each group was trying to get through all the portfolios for all the classes in a reasonable amount of time. Unless there was a disagreement about the grade, evaluators frequently did not see a reason to elaborate on their decisions; they were trying to move the process along.

But I do not believe this is the whole problem. I have seen similar results in group grading sessions where there was plenty of time; we were all going to be there until the meeting was over anyway. One evaluator says that the portfolio or the paper should pass because there are lots of good details and analysis, even though the organization is weak, and lots of people nod. But unless someone disagrees, the conversation rarely gets more specific.

This lack of discussion seems indicative of a larger problem. Because the process of assessment is messy and complex, and because we generally do not talk about what really value in the situations where we apply those values, we are driven to abstraction. "Focus," even "problems with focus," is an abstraction. Abstractions are not, in and of themselves, bad—they are actually quite useful. But they can be troublesome when they replace more thorough conversation in situations where complexity and thoroughness are necessary. In small group grading situations, using "live" student writing, such complexity may not be desirable, if only for the sake of the teachers' sanity as they try to finish grading in a reasonable amount of time. But complex conversation is certainly necessary, it seems to me, at the level of programs, whether that occurs in "norming" sessions, "articulation" sessions, or periodic faculty re-evaluations of grading criteria.

This is where I feel Broad's study (2003) provides a stronger and more systematic approach to articulating the substance of an assessment than any other method we currently have. City University's Dynamic Criteria Map is so complex that it requires a separate insert in the back of Broad's book. More important than the particulars of this map is the fact that each institution, each assessment—each scene, if you will—will produce a different map. There will likely be some overlap where programs value similar qualities; both Broad's and my own study include focus, organization, style, improvement or growth, and intellectual quality, though the language is not precisely the same. But as much significance lies in the differences: in my study, there were no decisional topics that covered whether students demonstrated "heart" or authority in their writing, just as there is no reference to the quality of research or use of source material in Broad's. These are places where the programs we examined differ, but they are also indications of differences in what we value when we read student writing. Of course we value different things; the context is different.

Although we both began with the idea of challenging accepted process—in my case by adopting trios and in Broad's by questioning norming and studying actual evaluative conversations[60]—we found, in addition, results that point at the substance of our assessments and how that substance differs from accepted standards in our programs. As with the theoretical principles that I have offered, it matters less if institutions adopt Broad's specific procedures for DCM than it does that we have seen the potential for examining the substance of our assessments beyond historically accepted norms. These two studies outline some of the possible results and suggest further work for programs and for researchers.

THEORY AND PRACTICE

The theory I am positing in this text helps us see the lack of discussion about primary substance in studies such as those at the University of Cincinnati and Washington State. The assessment process in both looks strong, both theoretically and practically, but work on substance specifically would more clearly indicate areas in need of additional work. The work in Broad's study and in Macoubrie's and mine gives us more information about what teachers actually value in student writing, and either DCM or decisional logics could help programs determine what substantive issues and changes they might explore. It is certainly likely that there are other methods for exploring this content, as well.

But more importantly, these last two studies in particular demonstrate the interconnectedness of meaningfulness and ethics. Broad would not have gotten the results he did if there had been no communal grading; without conversation, his method is all but impossible. The same is true for our study, and in that case, the limits of the conversation visibly limited our ability to understand how teachers made meaning during their evaluations. In both cases, silent and/or individual evaluation, even with reflection, would not likely have generated a sufficient level of detail and complexity to flesh out the substance of the evaluation.

Ethical assessment without attention to meaningfulness will tend to lead to assessment disconnected from context. At an extreme, this could mean a return to a focus on the instrument of assessment to the exclusion of the purpose and content or to the exclusion of disciplinary principles. More likely, it would result in an emphasis on the instrument or the process in ways that suggest, against the advice of assessment scholars, that instruments can be ported from one locality to another without careful consideration of the local scene. The process may be good and the participants may be treated well, but ethical assessment without meaningful assessment runs the risk of replicating recent large-scale assessment history, in which test-makers have looked for more and more reliable methods (in this case, ethical) without paying attention to the point of the test in the first place.

Meaningful assessment disconnected from ethics runs a different set of risks. Attention to content without concurrent attention to the process and the community is likely to result in assessments that resemble what we have had from ETS, if not in form at least in practice. In this scenario, meaningfulness would be determined by those in positions of power—WPAs, if we are lucky—without consideration for those in the classroom. This sounds a bit far-fetched, but how many of us have had our practices in first-year composition scrutinized and challenged by those in other disciplines who tend to disregard our expertise? In a perhaps kinder form, we may have this problem of meaningfulness without ethics under a standards system, particularly when those standards have remained the same, without reconsideration, for extended periods of time. Stagnant standards imply a kind of objectivity that is antithetical to the social constructionist principles that composition studies accepts. When those standards resemble the five developed by ETS in the early 1960s, the risk is greater.

This is not to say that an emphasis on one or the other is in and of itself counterproductive. Programs that have attended productively to the content of their assessments without looking at the ethics may well need to attend to the process for a while, and vice versa. But an emphasis to the exclusion of the other is potentially very damaging. We have been there before. And, in general, we are not satisfied with the results.

The work that has been going on during the last decade, as evidenced by the studies in this chapter, indicates changes in our approaches to assessing writing. Meaningfulness and ethics as a set of principles offers an alternative framework in which to evaluate those assessments—in which to understand the work we do when we evaluate student writing, particularly in large-scale assessment situations—and provides us with a variety of ways to develop alternatives more in keeping with our expertise, our disciplinary knowledge, and our values.