Even if by another name, writing assessment has always been at the center of work in writing; and it surely was there in 1950 when CCCC began. It wasn’t called assessment then, of course; that language came later. During the first of what I’ll identify as three waves in writing assessment, it was called testing, and it permeated the entire institution of composition—the contexts surrounding writing classes as well as the classes themselves, from the admission tests students completed to determine who would enroll in “sub-freshman English” and who in the “regular” course, to the grades that were awarded within the classroom, to the exit and proficiency tests that marked the single way out of many a composition program. Ironically, assessment in composition studies in those early days wasn’t just routine: it was ubiquitous—and invisible.

Like composition studies itself, however, writing assessment has changed during the past half century. One way to historicize those changes is to think of them as occurring in overlapping waves, with one wave feeding into another but without completely displacing waves that came before. The trends marked by these waves, then, are just that: trends that constitute a general forward movement, at least chronologically, but a movement that is composed of both kinds of waves, those that move forward, those that don’t. The metaphor of waves is useful conceptually, precisely because it allows us to mark past non-discrete patterns whose outlines and effects become clearer over time and upon reflection—and

Kathleen Blake Yancey is an associate professor at the University of North Carolina at Charlotte, where she teaches courses in writing, in rhetoric, and in the tutoring and teaching of writing. She has chaired the CCCC Assessment Committee, and has edited two volumes focused on assessment: Situating Portfolios and Assessing Writing Across the Curriculum. Yancey co-edits the journal, Assessing Writing, and her most recent writing project is the monograph, Reflection in the Writing Classroom (Utah State UP).
whose observation allows us in turn to think in an informed way about issues that might contribute to future waves.

An Overview

During the first wave (1950–1970), writing assessment took the form of objective tests; during the second (1970–1986), it took the form of the holistically scored essay; and during the current wave, the third (1986–present), it has taken the form of portfolio assessment and of programmatic assessment. This is the common history of writing assessment: the one located in method. But as Kenneth Burke suggests, other lenses permit other views; particularly when brought together, they allow us to understand differently and more fully. We could also historicize writing assessment, for instance, by thinking of it in terms of the twin defining concepts: validity and reliability. Seen through this conceptual lens, as Brian Huot suggests, writing assessment’s recent history is the story of back-and-forth shifts between these concepts, with first one dominating the field, then another, now both. A related approach constructs the history of writing assessment as the struggle between and among scholars and testing practitioners and faculty, those who speak the terms validity and reliability quite differently: the old expert; the new non-expert. From this perspective, the last 50 years of writing assessment can be narrativized as the teacher-layperson (often successfully) challenging the (psychometric) expert, developing and then applying both expertise and theory located not in psychometrics, but in rhetoric, in reading and hermeneutics, and, increasingly, in writing practice.

Still another way to trace the history of writing assessment is through its movement into the classroom; multiple choice tests standing outside of and apart from the classroom have become the portfolios composed within. And finally, writing assessment can be historicized through the lens of the self. Which self does any writing assessment permit? As important, given that “tests create that which they purport to measure” (Hanson 294), which self does an assessment construct? Portfolio assessment, with its multiple discourses and its reflective text, has highlighted this second question, certainly, but it’s a question to be put to any writing assessment seeking to serve education.

Significantly, these lenses don’t just frame the past; they point to the future, specifically to three issues. First, the role that the self should play in any assessment is a central concern for educators. It is the self that we want to teach, that we hope will learn, but that we are often loathe to evaluate. What is the role of the person/al in any writing assessment? A second future concern has to do with programmatic assessment: how can we use this kind of assessment—which is quite different than the individ-
ual assessment that has focused most of our attention for 50 years—to help students? A third concern focuses on what assessment activities can teach us: it's only recently that assessment was seen as a knowledge-making endeavor. Which raises a good question: what (else) might we learn from writing assessment? And how would we learn? Underlying these concerns is a particular construct of writing assessment itself: as rhetorical act that is both humane and ethical.

In itself, that understanding of writing assessment is perhaps the most significant change in the last 50 years.

A Context for a History of Writing Assessment: Spheres of Influence

During the first wave of writing assessment, dating from around 1950 to 1970, writing assessment was young, complex, conflicted. It was a critical time in that most of the issues that currently define the field were identified. Consequently, in our practices today, in the questions that continue to tease out some of our best thinking, we can trace the outlines of yesterday's concerns.

Much of the lore about the early days in writing assessment is accurate. It's true that "objective" tests, particularly multiple choice tests of usage, vocabulary and grammar, dominated practice. It's true that most testing concerns focused on sites ancillary to the classroom: typically, on the placement exercise used to "place" students into "appropriate" writing courses. And, in general, it's true that in the early days of CCCC, classrooms were defined, at least in part, by what we could call a technology of testing—not only by means of the tests that moved students in and out of classrooms, but also by way of contemporaneous efforts to bring "our work"—i.e., the reading and grading of student work—into line with testing theory. The early issues of CCC speak to these practices and aspirations convincingly: in summaries of CCCC workshops where the usefulness of objective tests is explained to the lay teacher, for instance, and in endorsements of national grading standards.

What's at least as interesting is a specific account of why these issues gained attention, why testing didn't include writing samples more often, why people who knew well how to read and value texts would turn to test theory when it came time to read and evaluate student texts. One contextual factor was demographic: the numbers and kinds of students that were beginning to attend school in the early 50's, students that seemed more and different than students we'd seen before. Consequently, there were genuine questions as to what to do with these students: where to put them, how and what to teach them, and how to be sure that they learned what they needed. In theory, those three tasks—(1) where to put students and (2) how and what to teach them and (3) how to be sure that they learned what they
needed—belonged to all educators. In practice, the tasks initially divided themselves into two clearly demarcated spheres of influence that characterize the first wave of writing assessment: the process of deciding what to teach the students belonged to educators, those who would become new compositionists; the process of moving students about, to testing specialists.

During the second two waves of writing assessment in composition studies—those of holistically scored essay tests, and next portfolios and program assessments—the two spheres merge and overlap, with administrators and then faculty taking on institutional and epistemological responsibilities for testing previously claimed by testing experts, and in the process developing a new expertise and a new discipline: writing assessment.

This history is also that story.

Methods and Sampling: Two Sides of the Same (Assessment) Coin

From the perspective of method, changes in writing assessment appear straightforward and familiar: from first-wave “objective” measures like multiple choice tests, largely of grammar and usage, to second-wave holistically scored essay tests to third-wave portfolios. Put another way, first wave evaluation relied on an “indirect” measure—a test of something assumed to be related to the behavior, but not the behavior itself (e.g., items like comma usage questions and pronoun reference corrections). Within twenty years, during the second wave, we began employing a “direct” measure—a sample of the behavior that we seek to examine, in this case a text that the student composes. Once the direct measure becomes accepted and even routinized as the measure of choice, the “one essay” model is soon replaced by a set of texts, so that: a single draft becomes two drafts; two drafts become two drafts accompanied by some authorial commentary; two drafts plus commentary become an undetermined number of multiple final drafts accompanied by “reflection,” and the set of texts becomes the new: portfolio assessment. As important as the method of assessing writing in this account is the sampling technique. The question “How shall we evaluate writing?” concerns itself not only with methodology, but also with behavior: which behavior should we examine? Sampling was critical, in part because sampling was (and is) the stuff of everyday classroom life: day in and day out, faculty assign, read, and evaluate student texts. In this sense, teaching writing is itself an exercise in direct measure. Accordingly, (1) teachers saw the difference between what they taught in their classes—writing—and what was evaluated—selection of homonyms and sentence completion exercises; (2) they thought that difference mattered; and (3) they continued to address this disjunction rhetorically, as though the testing enterprise could be altered—first on their own campuses; also at composition studies conferences like CCC, later and concurrently at testing-focused conferences like the
National Testing Network in Writing and the NCTE conferences on portfolio assessment; and concurrently in articles and books.

Still, it took over 20 years for this critique to make an impact, over 20 years for the second wave to occur. It’s fair to ask, then: if compositionists saw this disjunction between classroom practice and testing practice early on, why did it take over two decades to shift from one sampling technique to another, from one methodology to another? And the waves are overlapping, not discreet: why is it that even today, 50 years later, multiple choice tests continue to be routinely used in many assessment exercises (Murphy)²? The responses to these questions, phrased themselves as four general questions, are inter-related, each of them located in or deriving from the methods and sampling issues:

- What roles have validity and reliability played in writing assessment?
- Who is authorized and who has the appropriate expertise to make the best judgment about writing assessment issues?
- Who is best suited to orchestrate these questions, design an assessment based on the answers, and implement that design? In other words, who will wield this power?
- What, after all, is the overall purpose of writing assessment in an educational situation?

Each one of these questions points to one understanding of writing assessment; each one identifies a dimension of writing assessment still in contest.

Validity and Reliability: The Pendulum Swinging

Writing assessment is commonly understood as an exercise in balancing the twin concepts validity and reliability. Validity means that you are measuring what you intend to measure, reliability that you can measure it consistently. While both features are desirable in any writing evaluation, advocates of each tend to play them off against each other. Accordingly, which one should dominate, assuming only one could be favored, has generated considerable discussion—and change.

During the first wave, reliability prevailed; we see this, first, in the kinds of assessments that were commonly employed, and second, by the rationale for using them. That such tests were common is confirmed by various survey data. One survey in 1952, for example, included over 100 responding institutions and provided an all-too-common portrait: 90% of the responding institutions administered placement tests to entering freshman, 84% of those tests were standardized, and most of those tests were created
by external experts (Sasser 13). Similarly, the same survey reported that nearly half of the reporting institutions (44%) also included a test at the conclusion of the course, a “follow up re-test,” with half of the schools foiling that test score into the course grade. From placement to exit, then, objective testing defined the borders of many first-year composition courses, and in doing so, it also influenced what went on inside.

Such testing was theorized persuasively. Perhaps the most articulate theorist for this perspective was Paul Diederich, the ETS researcher whose *Measuring Growth in English* was the first monograph to address the specifics of postsecondary writing assessment. As an unofficial representative of the testing community, Diederich—at nearly every CCC during the 1950s’s and within the pages of *CCC*—repeatedly explained the value of the reliable measure, taking as his primary exemplar the prototypic placement test:

The best test to use at the college entrance level to pick out good, average, and poor writers is not a writing test at all but a long, unspeeded reading test. That will usually yield a correlation of about .65 with good teachers’ estimates of the writing ability of their own students in excellent private schools, in which a great deal of writing is required. Next best is a good objective test of writing ability; it will usually yield a correlation of about .60 with such judgments. A long way down from that is a single two-hour essay, even when it is graded twice independently by expert College Board readers. It will usually correlate .45 to .50 with such estimates. Furthermore, if you test the same students twice—in the junior and again in the senior year—the two reading tests will correlate about .85 with one another, while the two essays will correlate only about .45 with each other. Thus the reading test will not only pick out good and poor writers each year better than the essay but it will also pick out the same one both years, while the essay tends to pick out different ones. (qtd. in Valentine 90)

The logic here, admittedly, is compelling in its own way. If you want to predict how students will perform and if you want to do this in a fair, consistent, and efficient way, you go the route of objective testing—because even to generate the most favorable conditions with an essay, which of course are considerably less than those of the standardized test, you have to use expert College Board readers (an impossibility for the average campus), and even then the texts don’t correlate well with one another. Teachers’ estimates are just that: they cannot compete with correlations. Properly understood, then, the placement exercise is an exercise in numbers, not words.

Not that Diederich and other testing specialists didn’t consider the essay test (which, in fact, Diederich’s monograph both touts and complicates). It’s just that from both psychometric and administrative perspectives, the test-
ing task as they construct it is almost insurmountingly difficult because error-prone, inefficient, and expensive. You'd need, they say, "Six two-hour papers, each written in a separate testing session," read by "Four well-selected and trained readers, who will read and grade each paper independently," a process that costs "$100 per student" (qtd. in Valentine 91). What it all comes down to is twofold: (prohibitive) cost, of course, but also "the inevitable margin of testing error," a margin that is a given in a testing context, but that can be minimized. According to this view of placement (and testing more generally), what we need to do is to rely on "reliability of measurement"; "it is only when we use highly reliable tests that we come close to coming up with scores for an individual that are just measures of his ability—that don't seriously over-value him or under-value him" (Valentine 91).

The first wave of writing assessment is dominated by a single question: not the question we might expect—"What is the best or most valid measure of writing?"—but a question tied to testing theory, to institutional need, to cost, and ultimately to efficiency (Williamson)—"Which measure can do the best and fairest job of prediction with the least amount of work and the lowest cost?"

The answer: the reliable test.

The Discourse of a Writing Assessment: Tables Turned

But what about validity? This question, raised often enough by faculty, dominated the second wave of writing assessment. Faculty teaching in new open admissions schools and elsewhere saw new and other kinds of students; and an obvious discrepancy between what they did with their students in class and what students were then asked to do on tests (White Teaching and Assessing Writing). Their concern with validity was also motivated by the fact that by the 1970s, faculty had begun to identify themselves as compositionists. They knew more about writing: about writing process, about teaching writing process, about writing courses and what they might look like, about what composition studies might be. Given what we were learning, it made increasingly less sense to use tests whose chief virtues were reliability and efficiency. The shift to what did seem obvious—the essay test—had to be orchestrated, however, and it was, by two rhetorical moves, both of which worked inside psychometric concepts to alter assessment practice: first, to make validity (and not reliability) the testing feature of choice; and second, to undermine the concept of correlation as a criterion for evaluating tests.

Edward White took the first approach. As a faculty member who became an administrator—the first director of the California State University (CSU) Freshman English Equivalency Examination Program—he
understood the three variables that had to be accounted for in order to make essay testing feasible:

While some...chancellors, regents and the like are impervious to argument, most are not; many of those who employ multiple-choice tests as the only measure of writing ability are properly defensive of their stance but will include actual writing samples if they can be shown that writing tests can be properly constructed, reliably scored, and economically handled. (Teaching xiv, my italics)

Which is exactly what White and others—Richard Lloyd Jones, Karen Greenberg, Lee Odell and Charles Cooper, to name but a few—set out to do: devise a writing test that could meet the standard stipulated by the testing experts. To do that, they had to solve the reliability dilemma: they had to assure that essay tests would perform the same task as the objective tests. Administrators like White thus borrowed from the Advanced Placement Program at ETS their now-familiar "testing technology." Called holistic writing assessment, the AP assessment, unlike the ETS-driven placement tests, was a classroom-implemented curriculum culminating in a final essay test that met adequate psychometric reliability standards through several quite explicit procedures: (1) using writing "prompts" that directed students; (2) selecting "anchor" papers and scoring guides that directed teacher-readers who rated; and (3) devising methods of calculating "acceptable" agreement.3 By importing these procedures, test-makers like White could determine both what acceptable reliability for an essay test should be and, perhaps more important, how to get it.4 The AP testing technology, then, marks the second wave of writing assessment by making a more valid, classroom-like writing assessment possible.

At the same time that administrators and faculty were showing how a more valid measure could also meet an acceptable standard of reliability—and therefore how testing could be more congruent with classroom practice—other administrators and faculty were demonstrating in the language of testing why the reliable-only test was particularly incongruent. In 1978, for instance, Rexford Brown made this case not only by appealing to the context of assessment, but also by connecting that test to the context of the larger world:

Of course these [objective] tests correlate with writing ability and predict academic success; but the number of cars or television sets or bathrooms in one's family also correlate with this writing ability, and parental education is one of the best predictors there is. All existing objective tests of "writing" are very similar to I.Q. tests; even the very best of them test only reading, proofreading, editing, logic and guessing skills. They cannot distinguish between proofreading errors and process errors, reading problems and scribal stutter,
failure to consider audience or lack of interest in materials manufactured by someone else. (3)

The *correlations* here correlate with more than predictive ability: they are themselves a measure of affluence, of the *number of cars or television sets or bathrooms in one’s family*, and of another variable, *parental education*. Are these, Brown implicitly queries, the items we seek to test? Moreover, given the discrepancy between the items on the test and what we in our classrooms teach, what could such *scores* based on such items really *mean*, anyway? Meaning is, after all, located in more and other than correlations: it is intellectual and rhetorical substance.

By working both within and against the psychometric paradigm, then, faculty and administrators moved us during the second wave of writing assessment closer to classroom practice.

The Third Wave: New Assessment as Politics of Location

During the second wave of writing assessment, not all faculty, and not all institutions, were carried along: many of both continued the objective measures of the first wave, particularly when they engaged in placement assessments, and many continue these practices today. The first wave, in other words, hasn’t disappeared. And yet, at the same time, waves feed into other waves: just as the first wave fed into the second wave, the second wave itself began to make room for the third, again because classroom assumptions and practices could be translated into an assessment scheme. Put simply: if one text increases the validity of a test, how much more so two or three texts? In responding to this question, Gordon Brossell forecast the preferred technology of the third wave, the portfolio:

we know that for a valid test of writing performance, multiple writing samples written on different occasions and in various rhetorical modes are preferable to single samples drawn from an isolated writing instance. But given the sizable and growing populations of test takers and the increasing costs associated with administering tests to them, the problems of collecting and scoring multiple writing samples are formidable. Until we find ways to reduce testing costs and to improve the validity of the assessments, the whole enterprise is not likely to serve any purposes higher than routine sorting and certifying. (179)

As Writing Program Administrators, Peter Elbow and Pat Belanoff in the mid-1980s found a purpose *higher than routine sorting* when they directed a first-year composition program that—like the programs of the 1950s—required an exit exam. Dissatisfied with its form (it was a second-wave
essay test), Elbow and Belanoff used classroom materials to create a writing assessment embodying Brossell’s description: *multiple writing samples written on different occasions and in various rhetorical modes.* Or, a portfolio.

This model of writing assessment, with its different genres and multiple texts and classroom writing environment, seemed more valid still. But built into the model was another new feature, a reliability based not on statistics, but on reading and interpretation and negotiation. Rather than use an elaborated holistic scale (with a 1–6 scoring range, for instance), theirs required a simple dichotomous (if still holistic) decision: pass or fail. The “raters” were the classroom teachers themselves. They were not trained to agree, as in the holistic scoring model, but rather released to read, to negotiate among themselves, “hammering out an agreeable compromise” (Elbow, Portfolios xxi). Elbow called this a “communal assessment,” and argued that it was both more realistic and productive:

> the more we grade with others, the more attuned we become to community standards and the more likely we are to award grades fairly to all. Even though we know, for example, that we are passing a paper because of the quality of its language, we can become aware that the rest of the group would fail it for faulty thinking, and we can then recognize that all of us need to rethink and perhaps adjust our standards. And the greatest benefit of all comes when we return to our classrooms enriched by new ways of commenting on student texts which have come to us during discussions with our colleagues. (xxi)

In the late 1980s and into the 1990s, other portfolio systems developed, notably the portfolio-based basic writing program at Purdue University, the exemption program at Miami University, the placement program at the University of Michigan, and the rising junior portfolio at Washington State University. Simultaneously, individual faculty began using writing portfolios, sometimes as a means of formal assessment, sometimes as a way of learning. All of these portfolio assessments expressed a direct connection to classroom practice.

The Elbow-Belanoff model, however, was the first to raise several interesting, still-unresolved questions challenging both theory and practice in writing assessment. First, this model quite deliberately conflated two different processes—*grading,* typically a non-systematic classroom evaluative procedure, and the more psychometrically oriented operation called *scoring* (which involves the technology of scoring described above). Consequently, psychometric reliability isn’t entirely ignored, nor is its historical concern for fairness—Elbow and Belanoff stipulate that *award[ing] grades fairly to all* is a prime objective. But the mechanisms of classic reliability are supplanted
by different understandings and larger concerns: (1) about how raters can be led to agree (through “negotiation” rather than training); and (2) about the value of such agreement (desirable, but not required nor expected). Second, faculty are readers, not raters. As readers, they are guided rather than directed by anchor papers and scoring guides; and they are asked to read portfolios with the understanding that readers necessarily will value texts and textual features differently, that they will disagree, that they should negotiate. In sum, we see in this model a shift from a desire for the uniform replication of scoring practice to an assumed negotiation and acceptance of different readings. It’s only through the articulation of difference and negotiation, Elbow and Belanoff say, that community standards are developed, and through these standards that fairer grades can be derived. Moreover, they claim, this process enables us to refine responding skills that can be taken back to the classroom. This model of assessment, then, functions three ways: (1) as a sorting mechanism (pass-fail); (2) as a check on practice; (3) as a means of faculty development.6

It’s worth noting that this model of assessment—one that emphasizes validity at the same time it re-contextualizes reliability—emerged from a different context than the one primarily responsible for shaping earlier assessments. During the first and second waves of writing assessment, the common reference point against and from which reform occurred was the placement exercise, which is conducted as an extra-curricular exercise, one prior to college matriculation. By contrast, in early iterations of programmatic portfolio assessment, the initial reference point is curriculum-based, occurring (like the AP exams) at the end of a course—where it’s difficult to ignore the program you’ve just delivered, to bifurcate that program from a high-stakes assessment marking the students you’ve just taught in that program.7 Or: like other disciplines, writing assessment functions out of a politics of location.

Faculty experience with portfolios has raised three other, also-unresolved theoretical issues related to portfolios and to assessment more generally: (1) the nature of reading processes and their relationship to assessment; (2) the role of scoring procedures in an assessment; and (3) what writing assessments can teach us when they are located in practice.

Precisely because portfolios are “messy”—that is, they are composed of multiple kinds of texts, and different students compose quite different portfolios, even in the same setting and for the same purposes, which in turn can make evaluating them difficult—they invite several questions. Certainly, portfolios should be read: but how? Initially, it was assumed that faculty would read portfolios similarly (Hamp-Lyons and Condon), but this given has been contradicted experientially. More specifically, several recorded communal “readings” have suggested at least two quite different portfolio reading processes—one, a linear process; and a second,
"hypertextual"—and these different processes have called into question *how any text is read* in an assessment context unconstrained by the technology of holistic scoring (Allen et al., “Outside”). A second issue focuses on the propriety of portfolio scoring: like Elbow and Belanoff, others have questioned whether or not portfolios should be scored at all. A perhaps more interesting and related question has to do with whether a single holistic score is appropriate given the complexity of the materials being evaluated (e. g. Broad). A third issue inquires into the nature of collaborative scoring—a later version of communal scoring—and the value of including a mix of perspectives from various “stakeholders,” for instance, a set of outsider and insider readers, or a set of administrative, faculty, and external reviewers (Allen et al., “Outside,” *Situating;* Broad). A fourth issue focuses on what can be learned about our own practices from portfolios; exemplifying this aspect is Richard Larson’s review of curriculum as it is evidenced in a set of working folders.

Together, these concerns illustrate a new function identified for writing assessment during the third wave: creating knowledge about assessment, of course, but also about our own practices. When writing assessment is located within practice, its validity is enhanced, to be sure. But equally important, it reflects back to us that practice, the assumptions undergirding it, the discrepancy between what it is that we say we value and what we enact. It helps us understand, critique and enhance our own practice, in other words, because of its location—in practice—and because it makes that practice visible and thus accessible to change.

**Experts and Amateurs**

Another way of understanding writing assessment in the last 50 years is to observe that expertise in writing assessment has been redefined and created anew. During the first wave, testing specialists dominated the field, articulating in a testing jargon why (testing) things are. To create the second wave, a new, hybrid expertise developed: one influenced by testing concepts like validity and reliability, but one also influenced by pedagogical knowledge, composing process research, and rhetorical study. Simply put, we had a clearer sense of what we were doing in class, we began to administer programs, and so we began looking for ways to accommodate the assessment needs of our institutions to our own classroom practices. To do that, we, like Sylvia Holladay, began to develop our own expertise, an expertise located in two disciplines—*writing* and *assessment*. Likewise, we began to develop the disciplinary machinery that would support and disseminate such expertise: organizations like the National Testing Network in Writing; and books like Charles Cooper’s and Lee Odell’s NCTE edited collection
Evaluating Writing, the Faigley et al, Assessing Writers’ Knowledge and Processes of Composing, and the Williamson and Huot edited volume Validating Holistic Scoring for Writing Assessment.

In the third wave, another shift regarding expertise is underway; this one appears to be bimodal. On the one hand, as indicated in documents like the CCCC Bibliography, the CCCC Position Statement on Writing Assessment, and the CCCC program proposal forms, composition studies recognizes writing assessment as a field; we have a journal devoted exclusively to the discipline, Assessing Writing; and graduates of programs in rhetoric and composition contribute to their own programs as well as to the composition studies literature (books like New Directions in Portfolio Assessment and Assessing Writing across the Curriculum; articles in WPA: Writing Program Administration and the Journal of Teaching Writing as well as in CCC).

On the other hand, there still continues reluctance at best, and aversion at worst, to writing assessment. Sometimes it appears in resistance to grading practices (see Tchudi); sometimes it’s identified as the villain when a major educational disenfranchising event occurs (as with the current eradication of many basic writing programs); often it’s evidenced in a generalized faculty reluctance to take on the tasks entailed in any assessment. And sometimes, discomfort at least is articulated quite clearly when faculty practicing assessment presume to something quite different, a deliberately non-expert status, as Elbow and Belanoff suggest:

First, we note that we are not assessment specialists. We have not mastered the technical dimensions of psychometrics. That doesn’t mean we don’t respect the field; we agree with Ed White that one of the greatest needs is for practitioners and theorists like us to talk to psychometricians. But we don’t feel comfortable doing that so long as they continue to worship numbers as the bottom line. We think teaching is more important than assessment. (21)

Still associated with number-crunching and reductionism, assessment expertise is, at least sometimes, foiled against a teaching grounded in humanism. Ironically, it’s expertise rooted twice, in teaching knowledge and in assessment non-expertise; they seem to work together. At the same time, other new experts—theorists like Huot and theorist-practitioners like Michael Allen and Jeff Sommers and Gail Stygall—understand writing assessment itself as the grounds for that same humanism. They argue that the humanistic endeavor requires a student-informed and -informing assessment and the expertise that can create it.

Faculty experience with portfolios as an assessment technology has focused our attention from yet another perspective: that of practice. The effect of this practice has been to suggest new understandings about the kinds of expertise that might inform our assessment practices, with the
specific effects of democratizing, localizing, and grounding expertise of three kinds: student expertise, reader expertise, and theorist expertise.

- **First, student expertise.** Through the reflective texts in portfolios, students are asked to demonstrate a kind of expertise about their own work; their "secondary" reflective texts are used as confirming evidence of student achievement as documented in a primary text (Yancey, "Reflection"). Writing well is thus coming to mean twofold: writing well and being an expert on one’s writing.

- **Second, reader expertise.** Assessment specialists are looking more carefully at what they are calling “expert” readers, based on a second-wave holistic model that Bill Smith used at Pittsburgh and was later adapted for portfolio assessment by Washington State (Haswell and Wyche-Smith). In this model, readers are expert in a local sense— authoritative about the relationship between a student and a specific course, one that the teacher-reader has very recently taught. Conceived of this way, reliability is not a function of agreement, directed or otherwise, among raters so much as it is a function of rater experience with particular curricula.

- **Third, theoretical expertise** that grows out of and is integrated with practice. The practical work in assessment undertaken during the third wave has created a body of rich data permitting theories of writing assessment to emerge. The theories are developing in two ways: as elaborations and new applications of assessment theory generally (Huot); and as readings of practice suggest (Broad; Allen; Allen et al., “Outside”).

**Orchestrating Assessment: Politics of Location, Plural**

Closely related to the issue of expertise is that of power. During the first wave of writing assessment, faculty seemed content to allow testing specialists to direct the tests while they directed classroom activities: a kind of specialization of effort mimicked what appeared as a co-distribution of power. During the second wave of writing assessment, faculty began to see writing assessment as something that wasn't tangential to the classroom, but important in its own right, as Daniel Fader suggests: “…writing assessment is to be taken seriously because its first purpose is to determine quality of thought rather than precision of form. As our students, our readers, and our network of cooperating teachers have told us, it matters because it tries to test something that matters so much” (83). Assessment within the classroom thus took on increased emphasis and importance. Two examples—one focused on the role of error and another on response
to student texts—illustrate how assessment concerns begin to move inside the classroom, become transformed in that context, and generate new questions for assessment specialists and compositionists alike.

During the first wave of writing assessment, error (by means of test items) outside the classroom determines which classroom a student enters: error has an ontological status of its own. During the second wave, error comes inside the classroom: taken together, errors weave a pattern amenable to teacher observation and intervention (Shaughnessy). Still understood as mistakes, they become clues allowing a teacher to plot where to start and what to do. During the third wave, pattern of error is its own discourse: errors work together to create unconventional readings that evidence their own uncommon logic (Hull and Rose). Originally an external marker of deficit, error thus moves into the classroom and becomes its own legitimate text, a means of knowing for both student and teacher.

A similar kind of movement occurs with response to student writing. During the first wave of writing assessment, considerable comment is provided on how important response is in helping students: the early pages of CCC speak to this near-universal concern eloquently. But assessment as a discipline, located outside the classroom, includes no provision for response: it’s a null set. During the second wave, we see the first formal study of response, conducted by Nancy Sommers in 1981. Located not outside the classroom but inside, Sommers’ study is based in and oriented toward recommending good classroom practice. During the third wave of writing assessment, modes of response and their functions—when to praise, how to help students move toward reflective writing, and how students interpret our comments to them—have become a central concern (Daiker; Anson; Chandler). A current debate: should preferred response always be non-directive (Straub and Lunsford), or should it be situated (Smagorinisky)? As significant, response is theorized newly, not as an evaluative end, but rather as an inventive moment in composing. It’s a text in its own right, another place to continue what Joseph Harris calls the opportunity for writers to “change not only their phrasings but their minds when given a chance to talk about their work with other people” (68). Moving inward now—into the classroom and then into and within composing itself—writing assessment becomes social act.

As social act, writing assessment exerts enormous influence, both explicitly and implicitly, often in ways we, both faculty and students, do not fully appreciate. Certainly, writing assessment has been used historically to exclude entire groups of people: White makes the point that a primary motivation for holistic scoring was explicitly political, to enlarge and diversify the student body. Portfolios, for many, developed from similar impulses, as Catharine Lucas notes: portfolios provide for what she calls “reflective
evaluation,' a kind of formative feedback the learners give themselves” (2). Through this technology, then, “students’ own evaluative acuity is allowed to develop” (10). That is the hope. As others make clear, the hope is not yet realized. Two of the early portfolio models no longer exist: the Elbow Bel-anoff model is now defunct, a victim of politics; the University of Michigan portfolio program is rumored to be in demise along with its Composition Board; many other models oppress more than make possible, as Sandra Murphy and Barbara Grant detail. Beyond portfolio as technology, scholars continue to look, with depressingly frugal effect, for assessments more con-gruent with other epistemologies, like that of feminism (Holdstein); with other rhetorics, like that of African Americans (Ball); with other composing technologies, like that of hypertext (Yancey, “Portfolio”).

How these issues play out—and how we compositionists alleviate or ex-acerbate them—is a central and largely unexamined question for assess-ment, as Pamela Moss explains, one ideally suited to program assessment. This kind of assessment provides another lens through which to under-stand our practices and their effects, so that we might, ultimately and in a reflective way, take on the central question that doesn’t seem to surface of-ten enough: whose needs does this writing assessment serve? (Johnston) In detailing such a program assessment, Moss focuses on the power of naming and of forming that assessment wields: how, she asks, do students and oth-ers come to understand themselves as a result of our interpretations, our re-presentations, our assessments? How does such an interpretation impact students’ “access to material resources and [how does it] locate them within social relations of power” (119). Moss argues that in taking up these questions, it is insufficient merely to interview and survey students. “Rath-er, it is important to study the actual discourse that occurs around the prod-ucts and practices of testing—to see how those whose lives a testing program impacts are using the representations (interpretations) it produces” (119). Writing assessment here, then, is rhetorical: positioned as shaper of students and as means of understanding the effects of such shaping.

Writing Assessment and the Self: A Reflecting Lens

As Lester Faigley and James Berlin have suggested, education ultimately and always is about identity formation, and this is no less true for writing assessment than for any other discipline. What we are about, in a phrase, is formation of the self: and writing assessment, because it wields so much power, plays a crucial role in what self, or selves, will be permitted—in our classrooms; in our tests; ultimately, in our culture. The self also provides a lens through which we can look backward and forward at once, to inquire
as to how it was constructed during the three waves of writing assessment as well as how it may be constructed in the fourth.

During the first wave of writing assessment, the tested self of course took very narrow forms. In multiple choice tests, the self is a passive, forced-choice response to an external expert's understanding of language conventions. Agency is neither desired nor allowed. During the second wave, the self becomes a producer—of a holistically scored essay—and thus an agent who creates text. Still, there is less agency there than it appears. The text that is created is conventionally and substantively determined—some might say overdetermined—by an expert who constrains what is possible, by creating the prompt, designing the scoring guide used to evaluate the text, training the readers who do the scoring. Given these constraints, the authorship of such a text is likely to be a static, single-voiced self who can only anticipate and fulfill the expert's expectations, indeed whose task is to do just that (Sullivan). At best, agency is limited; a self-in-writing is permitted, but it is a very limited self, with very circumscribed agency. The text does not admit alternative discourses conceptually or pragmatically: it's text as correct answer.

During the third wave of writing assessment, the self emerges, and it's often multiple, created both through diverse texts and through the reflective text that accompanies those texts. And yet many are uncomfortable with this model of assessment, as Charles Schuster argues. He takes issue particularly with the reflective text since it invites portfolio readers to "fictionalize" authors:

In effect, fictionalizing student authors moves readers away from normed criteria, replacing careful evaluation with reader response.... Presumptions concerning personality, intention, behavior and the like skew readings or turn assessment into novel reading...Such fictionalizing serves a useful purpose within a classroom.... Writing assessment, however, demands that we exclusively evaluate what the student has produced on the page in the portfolio. Fictionalizing in this context can only obscure judgment. (319)

Others, however, aren't so sure, some pointing out that such fictionalizing occurs even within singular texts (Faigley; Sullivan), and others that such "narrativizing tendencies" are inevitable (Schultz, Durst, Roemer):

This narrativizing tendency constitutes one of our primary ways of understanding, one of our primary ways of making sense of the world, and is an essential strategy in comprehension. As far as portfolio evaluation is concerned, rather than say that narrativizing is right or wrong, perhaps we should start by admitting its inevitability, and by advising teachers to be aware of this tendency and not overvalue the stories we create. (130)
The questions raised within this portfolio assessment, then, take us back to reliability and link it to the personal: how do we read, particularly this kind of text; what do we reward when we read; and what (different) role(s) should the answers to these questions play in both classroom and external assessment? Or: where and when does the self belong, and why?

A final point: the self is constructed quite explicitly through reflection, it’s true. But the self is constructed as well through multiple school discourses—academic writing, writing connected to service learning, writing within disciplines, writing for the workplace, writing for the public. Each of these rhetorical tasks assumes a somewhat different self: how are these selves represented—or even evoked—in writing assessment? Or: how could they be represented in writing assessment, particularly one that is linked to democracy?

The Role of CCCC: Writing Assessment within Composition Studies

Through its conferences and within the pages of its journal, as the many citations here attest, the Conference on College Composition and Communication has provided the place where postsecondary writing assessment has developed as a discipline. Reading through 50 years of volumes impresses one with the sophistication of the issues raised, the commitment of compositionists to their students, the frustration that easier answers were not to be had. In addition to numerous articles of various kinds—theoretical, pedagogical and research—the pages of CCCC include writing assessment’s disciplinary consolidation, as we see in ever-longer, ever-more-complete and rhetorically informed bibliographies—in 1952, in 1979, and in 1992. It was likewise within the pages of CCCC where the first comprehensive statement about writing assessment was published, The Position Statement on Writing Assessment, which moves from what we know to include what that means: what we can and must do because of what we know. Because literacy is social, this statement claims, assessment must be specific, purposeful, contextual, ethical. And because it is social, we—students, faculty, administrators, legislators—all have rights and responsibilities.

Which is not to say that the story of writing assessment is a narrative of uninterrupted progress; it’s rather of narrative of incomplete and uncompleted waves: the early wave, governed by the objective measure; the second wave, which saw the move to the more valid holistically scored essay; the third wave, where portfolios contextualized our students’ work and invited us to consider how we read, how we interpret, how we evaluate. At the same time, energies are currently accumulating as though to gather a fourth wave. Perhaps this next wave will focus on program assessment as epistemological and ideological work; perhaps it will focus more on
individual assessment as interpretive act; perhaps it will take on the challenges posed by non-canonical texts (email, hypertext, multi-generic writings); perhaps it will address the kinds of expertise and ways that they can construct and be represented in writing assessment; perhaps it will include topics that are only now forming.

What is certain is that writing assessment is now construed as a rhetorical act: as its own agent with responsibilities to all its participants. What’s also certain is that practice has located assessment, even during the first wave, and that practice has motivated successive waves of development, permitting a kind of theorizing that is congruent with a composition studies that is art and practice. Grounded in such practice, then, writing assessment is becoming more reflective about that practice and the practices to which it connects, uncovering assumptions we bring to texts and that we create as we read texts, understanding our work in the light of what others do, apprehending that what we do is social and thus entails both ideological and ethical dimensions that themselves are increasingly very much a part of both theory and practice.

Acknowledgments: Thanks to Russell Durst, to Cynthia Lewiecki-Wilson, and to a third anonymous reviewer for their clarifying and productive remarks on this essay.

Notes

1. Of course, how and what to teach them—that is, what the content of the English course should be—was also a frequent topic in the early days.

2. Even as I write this, there is a call on the listserv WPA-L inquiring into how multiple choice tests can assist in placement.

3. As I suggest later in this paper, the fact that this is a classroom based program is significant.

4. And White demonstrates: “the actual reliability of the essay test therefore lies between the lower limit of .68 and the upper limit of .89; in practice, this means that any two readers of the test, working independently on a six-point scale, agreed (or differed by no more than one point) approximately 95 percent of the time.” (Teaching 27)

5. Ironically, what Brown recommends seems considerably less progressive: “Use computers. Have people mark off T-units in the essays so you can gather information about number of words per T-unit, number of clauses per T-unit, Number of words per clause, number of adjective clauses, number of noun clauses, and so on—information about embedding, in short, which ties you directly to indices of syntactic maturity” (5).

6. Assessment—or the specter thereof—has sparked many a faculty development program. See, for instance, Toby Fulwiler and Art Young’s account of the way the WAC program at Michigan Tech began, in their introduction to Assessing Writing across the Curriculum.

7. The first books on writing portfolios all concern themselves with portfolio practice as it occurs in classrooms or just after. See Belanoff and Dickson, and Yancey, Portfolios.

8. Reflection is increasingly a part of non-portfolio placement exercises: at Coe College, for instance, at Morehead State, and at Grand Valley State College. And these practices show how waves overlap in still other ways, here in a second-wave essay format enriched by a third-wave reflective text. For more on self-assessment in placement, see Royer and Gilles’ “Directed Self-Placement.”
Works Cited


Fader, Daniel. “Writing Samples and Virtues.” Greenberg et al. 79–92.


