

*H*olistic scoring, as Edward White notes in a historical overview of changes in assessment practices since the 1970s, was introduced in the 1970s as "a flexible, accurate, and responsive measurement method, one that could come under the control of teachers" who were dissatisfied with indirect measures of writing ability (*Teaching* 270). For many years, holistic scoring has been perhaps the primary scoring method associated with large-scale direct writing assessments such as placement tests. But as White notes in his retrospective, holistic scoring has properly become the subject of scholarly analysis, and developments in assessment theory now challenge traditional methods of holistic scoring. These developments offer writing program administrators new options for the design of scoring sessions, and suggest that a tight-knit community of teachers can maintain sufficient agreement about the requirements for success in an institution's writing program. In the past few years, administrators designing scoring for a direct writing assessment would have considered holistic scoring the only logical option; now, administrators have a variety of models to examine.

The newer models rely on teacher expertise to sort students into the appropriate courses.¹ One model, developed at the University of Pittsburgh by William Smith, posits that placements are best decided by teachers of particular courses; another model, developed at Washington State University (WSU) by Richard Haswell and Susan Wyche-Smith, posits that qualified raters can easily identify prototypical placements into first-year composition courses, and that placements into other courses, as well as marginal first-year composition placements, require the time and attention

New Visions of Authority in Placement Test Rating

Susanmarie Harrington

of a group of expert raters. Both these models challenge the holistic scoring assumptions that have guided a generation of direct writing assessments, and it is the purpose of this study to test the assumptions in such teacher-driven placement tests. What factors drive good placement decisions? What contributes to decision-making expertise in such a system? In order to answer such questions, I designed an empirical study that tested the impact of placement test rater meetings on rates of agreement and placement outcomes. Such a study, based on a summer's worth of placement tests at a large urban university, provides a rich array of data for the testing of theoretical assumptions about placement models. Given that placement testing is practiced at most American universities and that a timed impromptu is the most common form of placement tests (Murphy et al.; Huot, "Survey"), such inquiry has wide applicability.²

Before describing the forces on my campus which made such study a matter of practical as well as theoretical utility, I will briefly review the foundations of holistic scoring and the challenges raised by the WSU and Pittsburgh systems.

Challenges to Holistic Placement Scoring

Holistic and primary trait scoring sessions require raters to use a central scoring guide in their decisions. A traditional holistic scoring rubric provides descriptive paragraphs corresponding to each point on the rating scale; most scales use 4 or 6 points. A primary trait rubric provides descriptive paragraphs for each trait or dimension of text scored (such as style or organization; see Lloyd-Jones for more information). Central scoring guides provide clear parameters for raters, and they impose order on the messy business of evaluating student writing (see White, *Teaching* ch. 10; or Cooper). The scoring guide is the key to the reliable assessment of texts, for it enables all raters to work from the same foundation. Without such a foundation, raters would apply individual criteria, leading to idiosyncratic and unreliable scoring patterns. A well-run holistic scoring program provides, as White explains, "quick, economical, and reasonably reliable rankings of large numbers of test papers" ("Apologia" 31). Experienced holistic scorers achieve impressive degrees of inter-rater reliability, which promotes confi-

dence in test results.

The reliability of a good holistic scoring system is maintained by thorough training and rating sessions that open each rating session (and may be repeated throughout the day, depending on the assessment leader's design). Anchor texts, chosen because they exemplify the rubric's description of various levels of achievement, are used to ensure that all raters participating in the day's work agree on rating standards. The assessment leader's job is to monitor rating and ensure that all participants are understanding and implementing the centralized rubric. Discussions of the anchor papers translate the scoring guide into reality. As White notes, the discussion allows "readers [to] internalize and come to 'own' the scoring guide" under the watchful eye of the assessment leader (*Teaching* 203). A well-run holistic scoring session achieves reliable scoring by trained raters, allowing the evaluation of students' texts to proceed, no matter what differences in teaching styles, philosophies, or standards might be manifest among the raters in a less controlled situation.

However, one problem arises in the use of scoring guide-driven sessions when the test raters are writing teachers involved in making placement decisions. Traditional holistic scoring practices tend to suppress the connection between courses (and teachers) and scoring. Even though scoring rubrics are derived from the themes emphasized in the local curriculum, and are revised in light of actual student performance over time, they are not written in the language used to describe students in the classroom, and usually do not make explicit references to placements. Raters may be aware that if two raters give an essay a 4 on a 6 point scale, the result will be a first-year composition placement, but the scoring guide itself makes no mention of that. Rather, the focus of the scoring guide is the *text*, and the guide describes textual features at various levels of performance.

But teachers like to talk about students, not simply texts. Inevitably, during a placement rating discussion, raters begin to talk as teachers, rather than as users of a scoring guide, and begin to make decisions based on their classroom experience. One of the key jobs of a scoring session leader is to make sure that the raters adhere to the scoring guide, and the leader is constantly monitoring the conversation to ensure that the central scoring guide

remains the anchor for all decisions made. Nonetheless, conversations about holistic scores invariably veer into some remarks about students in class. "This student reminds me of the kind of student who sits in the back of the room and doesn't talk," raters will say, or "This student is the kind who just needs a conference on paragraphing and then the essay would be revised with no problems." At that point the session leader steps in with a comment like "Don't talk about the writer! Talk about the text and how it relates to our rubric." While this direction is successful in keeping the conversation related to the rubric, it is not successful in keeping teacherly thoughts out of raters' minds. Teaching experience is probably always a factor in the use of holistic scoring, although it is difficult, if not impossible, to study the relationship between overt attention to rubrics and perhaps unconscious reliance on prior experience in raters' use of the rubrics (see Barritt, Stock, and Clark for discussion or related issues; Broad discusses the relationship between teacher experience and scoring guidelines in a portfolio program). Yet in holistic scoring, there is no way to take account of teaching and classroom-based expertise.

For placement, then, traditional holistic scoring sessions involve a tension between teaching and scoring, or what Huot has called tension "between a reader as reader and a reader as rater" ("Literature" 255). This tension is converted into a strength in the Pittsburgh and Washington State University placement systems. These practices differ from the traditional holistically scored direct assessment in that they invite test raters to make direct decisions about placement, rather than indirect ones. Instead of considering whether a given essay is a 4 or a 3, raters can consider whether an essay most resembles texts produced in the early weeks of basic writing or first-year composition. These direct placement readings rely not on a rubric, but the local curriculum to define the different placement points. Teachers' and administrators' understandings of course goals and expectations of students at the start of the semester are used to analyze the possible responses to the placement test, although these understandings are not codified in a central document to which all raters must refer. These new procedures profoundly challenge the assumptions which have guided holistic scoring and thus may act as useful prototypes of assessment practices that implement

new theories of assessment attentive to, in Huot's formulation, "the context of the texts being read, the position of the readers, and the local, practical standards teachers and other stakeholder establish for written communication" ("Toward" 561).

The Pittsburgh and WSU Systems

At the University of Pittsburgh, William Smith developed a model he called "placement rating" (148).³ Placement rating differs from other scoring systems in that its purpose "is to use the student's text as a window into that student so as to place the student into the course which best matches his/her needs and abilities" (148). The rating scale is not keyed to some externally derived rubric, as in traditional holistic scoring, but to the available course options for an incoming student; furthermore, scores are not numeric, subject to adding and averaging. Test raters make direct decisions about what course a student needs to take, and reading procedures allow for continued reading until reader agreement is reached. Finally, the assessment keeps in mind that there will be "a very direct impact" on the students taking the test; "any error in placement will mean that the students are not being well served" (150). In holistic scoring, one scores the test; in placement rating, one places the writer. Advocates of holistic scoring, I must note, caution against the misuse of test scores, which can damage students, and note that test raters must feel a real sense of community, as well as ties to both scoring guide and examination, in order for holistic scoring to be effective (see White, "Holistic" 93). But the sense of community that forms a holistic scoring session revolves around the scoring guide, rather than a shared sense of teaching expertise, and this difference has great theoretical implications.

The heart of the shift from scoring the test to placing the writer comes in the location of *authority* in the assessment. Smith found that the "raters' expertise—the expertise which comes from working with their students—might be more powerful than any training session in which they are told about the various courses and read essays prototypic of those courses" (175); equally important, he noted that raters' experience can never be trained out of them. That is, leaders of holistic scoring sessions who ask

raters to disregard experience-based reasons for scores and adhere to scoring guide-based reasons for scores probably succeed in stopping the articulation of such reasons, rather than the use of them. Smith found that raters did the best job placing students when they accepted them into their own courses. However, "all raters, regardless of course-taught expertise, [were] able to reliably discern student who are prototypic of a course" (181). Smith also found that "the raters were highly reliable. They knew whether an essay fit into their course, and...they knew when it didn't" (185). Because of the effect of "course-taught" expertise, the only decisions that raters make in Smith's model are acceptances. Raters accept or reject students for the courses that they teach; if a reader rejects a student, the test is read by a reader who teaches an adjacent course, until the student is accepted into a writing course.

The implications of Smith's model—that teacher expertise is the source of authority for placement decisions—are embraced in a placement model designed at Washington State University (WSU) by Susan Wyche-Smith and Richard Haswell. While their model is very different in form from the Pittsburgh model, the fundamental assumption is the same: teachers know best which students belong in their classes. The Washington State system relies on the notion of *prototypes*, or texts that are so clearly in the bounds of a given course that raters with the proper expertise (course teachers) can easily recognize them. At WSU, a first tier of raters makes only one decision: does this student belong in first-year composition? If the answer is yes, no further reading is necessary. If the answer is no, a second tier of more expert raters comes in and makes a placement decision; these raters have more experience with the courses other than first-year composition. These expert raters may place some students into first-year composition (students whose writing falls on or near the boundaries between first-year composition and basic writing, or first-year composition and honors, for instance); their job is to make decisions about students whose texts are not prototypic of first-year composition.

Because both these models rely on the same assumption of teacher expertise, they avoid the previously-described problem which has plagued anyone who has run a holistic scoring session for placement tests: when test

raters are teachers, they want to use their teacherly expertise. These placement models capitalize on teacher expertise and make it central to the decision-making. The models have one key difference, however: the fundamental move of the scoring system. At Pittsburgh, raters of each test ask the question "Does this student belong in *my* course?" and the tests flow from rater to rater according to the way the first reader answers the question. At Washington State, raters of each text ask "Does this student belong in *first-year composition*?" and second raters are invoked only if the answer is no. Smith and Haswell and Wyche-Smith are careful to note that both their models were designed to meet local needs and should not be regarded as models to be copied. Placement testing involves a *local* decision; it determines which particular course a particular student needs to take. Decisions about placement testing need to take into account the nature of the student body, the nature of the courses, and the personnel available to make the placement decisions, in addition to concerns about the nature of the test itself. Although placement is always local, the similarities between the Pittsburgh and WSU systems can function as useful guides to others designing placement tests; thus it is important to see whether the assumptions that underlie these systems can function in other settings in ways that provide reliable and fair assessments for students.

In particular, it is useful to determine whether or not teacher expertise seems to have any bearing on the adequacy of placement decisions. Related to the issue of teacher authority or expertise in placement decisions is the issue of how authority and expertise are maintained over time. In traditional scoring models, training (via meetings) is the key to maintaining authority and consistency. And authority and consistency are closely related to two foundational testing concepts: *reliability* and *validity*. Any good assessment must have both these properties. A reliable assessment is a fair assessment, one which will consistently produce, for the most part, the same results. A valid assessment is one which assesses what it sets out to assess (in this case, students' ability to write in relation to the local curriculum divisions). Holistically scored placement tests are reliable assessments when experienced raters are well trained; the extent to which these new placement systems produce reliable results is critical. If teacher expertise leads to

reliability, then administrators and students can trust the resulting placements; if it does not, the placements fail to serve students' interests.

Daily rater meetings are the hallmark of holistic scoring sessions, and the foundation for scoring reliability. As Smith notes, teachers enjoy the chance to meet and discuss student writing. However, if teacher experience, rather than training in test rating, is the driving force in the assessment system, regular meetings should not be necessary. In a well-run writing program, faculty development opportunities should keep teachers aware of the goals and outcomes of each introductory level course. Some differences between teachers and between sections may exist, but a good deal of overlap should be created and shared. This sense of course boundaries is imparted by the regular formal and informal activities of a writing program, not by special scoring training. The implied argument in the Pittsburgh and WSU systems is that a vibrant writing program provides teachers with the training they need to make good placement decisions. And if teaching experience guides good placement decisions, raters must be teaching courses, but they need not be attending regular meetings. Can a well-run faculty development program create the shared standards that lead to reliable judgements? Another area of inquiry should involve whether raters can reliably identify prototypic essays. The main difference between the Pittsburgh and Washington State systems involves different valuing of course-taught expertise versus the ability to identify prototypical first-year composition placements.

A Program-Based Inquiry into Theoretical Questions

Context and Personnel

Recent changes in my department's placement testing situation provided a practical laboratory for the investigation of such questions. The logistics of our placement testing are dictated by campus priorities. The campus—an urban commuter campus with wide admissions standards—is currently reforming its application/registration/orientation process in an effort to reduce the number of visits prospective students must make to campus before classes start. The move to daily placement tests is part of the efficiency reforms, an easy enough change to make with the cooperation of the Testing Center. Previously, the department had offered tests only on certain

days each semester, and large batches of tests were then scored by test raters using holistic methods. But once we changed to daily testing, it became difficult to use holistic scoring methods, dependent on rater meetings. Because our test raters are all members of our part-time writing faculty, their teaching schedules are varied, and often quite tight. Trying to gather raters for meetings one day a week was practical at some points in the semester, but gathering test raters for daily meetings would be prohibitive. So the new placement methods that relied more on teaching experience and less on standardizing meetings seemed attractive for practical reasons, not just theoretical ones.

Currently, placement testing occurs six days a week, fifty weeks a year. Placement tests are rated each week day by eight members of the English department's associate (part-time) faculty; all have taught either first-year composition or basic writing for at least four years. Most have taught both courses, and four also work in the University Writing Center. At the time of this study, no rater had any experience teaching honors courses, an artifact of the way the department distributes small benefits to faculty of different rank. Placement rating is considered a "perk" for experienced part-time faculty, while honors teaching assignments are considered a perk for lecturers, whose heavy administrative loads do not allow time for placement test reading. This disjunction between the range of possible placements generated by the raters and their teaching experience leads to some problems, discussed below.

As I designed a new scoring procedure, I wondered whether it would work. How could we tell that raters were making good placement decisions? Would a system that largely abandoned standardizing meetings really work? With these questions in mind, we implemented the daily rating system. The results of the work described here led to changes in our scoring practices; the changes that were identified after data analysis are described below.

How Tests Were Scored

During the period of this study, tests were rated using a model inspired by the Pittsburgh and WSU systems, modified to meet local needs: two raters scored each test, and made a direct placement decision (e.g. "take first-year composition"). At the time these data were collected, the

raters could decide that a student needed to take honors composition, first-year composition, basic writing, or a pre-basic writing course (the pre-basic writing placements have been excluded from the discussion because they are rare). If raters felt that a test fell between two courses, they indicated that with an "in-between" rating that nonetheless showed which course the reader would lean towards for placement. Third raters were used only if the first two raters disagreed about what course a student should take.

Data Collection

Data were collected over a period of twenty-two weeks, during which 2,877 exams were given. Raters' meetings were varied in order to explore the effect of those meetings on placement decisions (the research design is graphically depicted in Figure 1). For the first six weeks of the study, raters met before each placement rating session, read 3 placement essays, and discussed placements. This sort of meeting, derived from the holistic scoring tradition, was designed to help raters articulate their notions of course boundaries and to feel "in sync" with each other; each reading session began with a common discussion of raters' reasons for making placement decisions into particular courses (such as basic writing or first-year composition). During this six-week period, the raters made their placement decisions independently, and had the option of negotiating differences in placement or sending disagreements on for a third read by another reader.

In the second six-week session (weeks 7-13 of the study), raters did not meet at all to discuss placement tests; they came in daily to read tests but did not discuss any tests with other raters. This (non)meeting period was designed to explore the assumption that if teaching experience alone drives the system's reliability, the absence of meetings should have no impact on placement decisions. Alternatively, if meetings did matter, rates of agreement would fall off over time. During this period, any disagreements about placement were passed on to third raters for a decision.

In the final weeks of the study (weeks 14-22), weekly placement meetings were held. All raters gathered on Fridays to read tests and discuss placements, after which the raters assigned to finish that day's test batch would do their work; on other weekdays, raters came in, rated exams, but did not meet with other raters before commencing their rating. This session,

which lasted for eight weeks (in order to make the number of tests read in this period roughly equivalent to that read in the other two sessions), was a return to the usual practice on our campus.

	Frequency of Rater Meetings	Rating Process	Result of Disagreement between Rater 1 and Rater 2
Phase 1	Daily	Two raters met each morning to discuss a group of placement tests and compare rating decisions; discussion focused on the rationale for ratings and identification (but not always resolution) of differences in opinions. Each rater then read the day's tests individually, making rating decisions independently of the other rater.	Test sent to 3rd Rater, or disagreement negotiated between R1 and R2
Phase 2	None	Two raters came to campus daily to rate the day's placement tests, but did not discuss rating with each other. Rating decisions were made independently. No rater meetings of any kind were held during this period.	Test sent to 3rd Rater
Phase 3	Weekly	All raters came to campus weekly to discuss placement tests and compare ratings; discussion functioned as in Phase 1, but in a whole group. Two raters came to campus each day during the week to independently rate each day's batch of placement tests.	Test sent to 3rd Rater

Figure 1: Research Design

In all three periods, at least two raters read each test, and each reading was independent. Neither reader was aware of the other raters' placement decision until the scores were recorded at the end. Raters signed up for reading days based on their own schedules; over the course of the study, reader pairs shifted, so that most raters read with each other over time. Test scores were entered into a database, along with information about reader teaching experience and years of test rating experience.

Results

Overall Rater Agreement and Soundness of Placement Decisions

Of the 2877 tests included in this analysis, 717 (25%) were placed

into basic writing; 1994 (70%) into first year composition, and 139 (5%) into honors. (A small number of tests were excluded from analysis, including placements into pre-basic writing or ESL classes as well as unscorable tests and exams with missing data.) The distribution of placements during the twenty-two week study period mirrored the distribution of placements more generally. During the period of this study, raters agreed with each other much more often than not about course placement. The overall rate of agreement about course placement was 82%; 18% of the tests required a third reader to make a placement decision.

Of course, agreement about placement does not, in itself, mean that the decisions are good decisions; raters could be agreeing about poor placements. In order to determine whether these placements were good ones, I surveyed writing faculty in the early weeks of the semester, and asked them to report any students in their sections who could have been placed into a higher or lower class; I analyzed overall course grades; and I analyzed internal rosters which listed each student's grade and an explanation for each non-passing grade. We have used these methods of establishing what Smith calls the adequacy of placement decisions since 1994; since the tests analyzed here were taken during the spring and summer of 1996, the placement adequacy studies done during Fall 1996 would have involved almost all of the students who tested during this study.

Following Smith's method, I asked teachers to identify which students they considered placed too high, too low, or just right in the third and fourth weeks of the semester (before that point, teachers may not be able to form an accurate judgment; past that point, the effects of instruction and teacher effort are such that familiarity leads teachers to see almost all students as appropriately placed). Between Fall 1994 and Fall 1996, the results of this faculty survey indicated that most students were, from their teachers' perspectives, in the right course. In more than 60 sections of first-year composition (nearly 1800 students per semester), an average of only 6 students were reported placed too high, and an average of only 12 students were reported placed too low (and those students were usually those who had taken basic writing more than once). The basic writing teachers' responses, however, indicated slightly more dissatisfaction with the placement results.

We offered about 22 sections of basic writing each semester (nearly 650 students per semester) and in Fall 1994 and 1996 there were roughly 20 students reported placed too low, nearly one in every section. In the spring 1995 and 1996 surveys, roughly one-third of sections reported one misplaced student. Smith's discussion of course boundaries suggests that some students will always be "between" courses; since students writing abilities do not neatly match curricular boundaries, the relationship between preparedness and curriculum is fluid. Nonetheless, the basic writing course coordinator felt it would be advantageous to try to make changes in the placement model that would direct a slightly larger number of students into first year composition; the changes that we identified are discussed later in this essay.

In order to analyze the reasons for student failure, in the first full semester after the data were collected, I also analyzed internal rosters on which instructors recorded reasons for students' failures. The overwhelming reason for failure to pass both the basic writing and first-year composition course was attendance. Very few students failed because their written work was not up to standard; rather, they failed because they stopped coming to class part-way through the term and never handed in any work.

Relationships Between Courses and Placement Decisions

While the overall rate of agreement about course placement during the course of the study was a respectable 82%, this rate shifted quite considerably when controlled for factors such as writing course, meeting frequency, and teaching experience, moving from a high of 93% agreement to a low of 58%.

Placement Decisions by Course

As Table 1 illustrates, the rate at which the initial two raters agreed with each other about a placement decision shifted significantly across course. It was much easier for raters to agree about placements into first-year composition than it is to make decisions about placements into any other course. Rater agreement was highest concerning placements into first-year composition: 86% of the time no third rater was needed. Raters agreed only 73% of the time with respect to placements into basic writing classes, and only 63% of the time about honors placements.

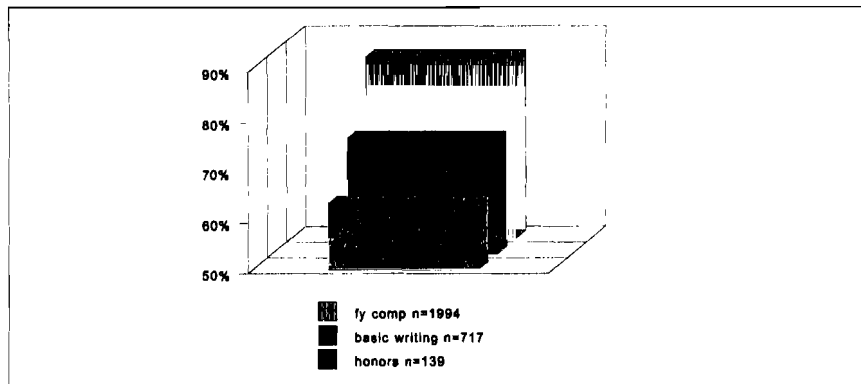


TABLE 1 - Rates of agreement by course

Meeting Frequency

Table 2 shows the placement agreement rates for the three courses, when the frequency of rater meetings is taken into account. Agreement about placements into first-year composition remained remarkably stable (hovering around 87%), no matter how (in)requently raters discussed tests together. Meeting frequency, however, did affect rater agreement about placement into basic writing and honors courses. Rater agreement with respect to basic writing placements was highest (80%) during the period with daily meetings; dropped down to 72% when meetings were not held; and dropped even further (to 58%) when weekly meetings were reinstated ($p \leq .001$ for all group comparisons). Rates of agreement regarding placement into honors also varied across time period, although not so dramatically as with the basic composition placements. (Variations here were not statistically significant, largely because of the smaller number of students placed into honors classes.)

Course Most Recently Taught

Smith's placement rating model is predicated on the notion that the best decisions are made by teachers of the course, and he found that he could account for almost all reader disagreement in this fashion. To see whether Smith's model accounted for disagreements on my campus, I measured the percentage of time the first rater's placement decision was subsequently confirmed—either by the second rater, or by the third. Table 3 shows the percentage of time Rater 1's decisions were confirmed, factoring in courses most recently taught. The effect of courses taught on placement

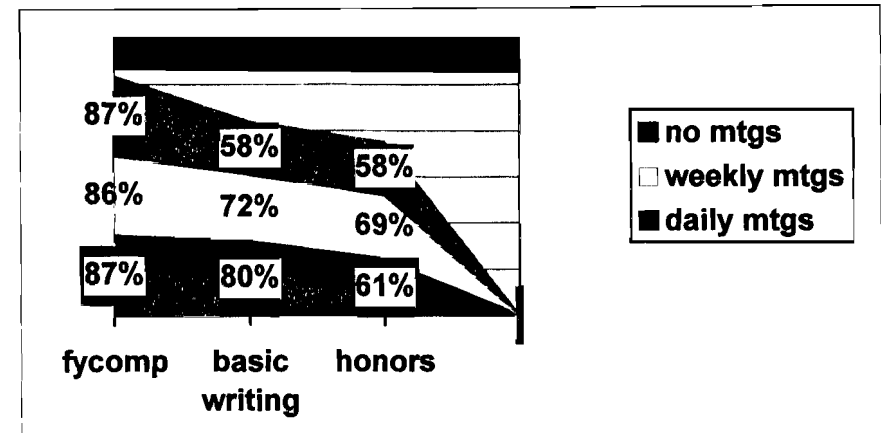


TABLE 2 - Rates of agreement, by course and by rater meeting frequency

into honors classes is not examined here, for the simple reason that the raters do not teach the honors sections.

As a general rule, the first rater's placement decision was confirmed 93% of the time when the placement was into first-year composition, and 87% of the time when the placement was into basic writing. When agreement with Rater 1 is measured against teacher experience, first year composition remains unaffected. Even when Rater 1 was not currently teaching first year composition, or had not been teaching first year composition in the previous semester, placements were confirmed by subsequent raters 92% of the time. Placements into basic writing courses, conversely, were affected by course most recently taught ($p \leq .01$). Counter-intuitively, those raters who had not just taught basic writing were more likely to have their place-

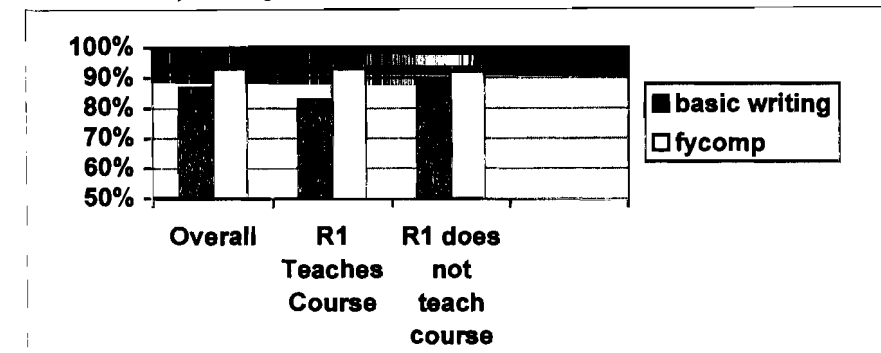


TABLE 3 - First rater teaching experience and rates of agreement, by course

ments confirmed by later raters (90%) than those who currently taught basic writing (83%).

Multivariate Analysis of Rater (Dis)Agreement

An obvious question at this point concerns the relative importance of the nature of the placement decision, meeting frequency, course most recently taught, and years of rating experience on placement decisions. The seeming impact of course taught on basic writing placement decisions, for example, may disappear when other relevant factors are taken into consideration. It's possible, for instance, that the real determinant of rater agreement about placements was simply years of rating experience. Perhaps raters who had worked the longest made the best decisions. Or perhaps raters were most likely to agree about placements they viewed as prototypical placements and less likely to agree about placements on the borders of the courses in question. In order to tease out the effects of all these factors, I regressed rater disagreement against four sets of variables: course most recently taught; rater experience; placement decision; and prototypicality of the placement decision.

The prototypicality of the placement decision was determined by whether or not the raters marked a placement decision as falling squarely within the bounds of a given course, or whether their decision indicated the placement in the boundary area that is arguably shared by two courses. Rater experience was coded as low, medium, or high, with low experience being one year, and high experience being more than three years of experience.

As Table 4 shows, the factors that exerted significant influence on rater disagreements about placement decisions are the course in question, the prototypicality of the student exam, and years of rater experience. When other variables were controlled for, neither meeting frequency nor course most recently taught exerted any significant influence on the rate of rater disagreements.

TABLE 4: Factors Influencing Rater Disagreement in Placement Decisions⁴

Course Placement		
Basic Writing (0-1)	.14***	(.02)
Honors (0-1)	.22***	(.03)
Prototypic (0-1)	-.10***	(.02)

Meeting Frequency		
Daily (0-1)	-.02	(.02)
Weekly (0-1)	.03	(.02)
Course Most Recently Taught		
First Year Composition (0-1)	.002	(.02)
Basic Writing (0-1)	-.05	(.03)
Rating Experience		
Rater 1 (0-2)	-.04***	(.01)
Rater 2 (0-2)	-.02*	(.01)
Constant	.29***	(.03)
F	16.42	
Number of tests	2,842	

Entries are unstandardized regression coefficients, with standard errors in parentheses. * $p \leq .05$ *** $p \leq .001$

The course in question had the greatest impact on raters' tendency to disagree about placement. For instance, when other factors were held constant, Raters 1 and 2 were 14% more likely to disagree about placements into basic writing than placements into first-year composition, and 22% more likely to disagree about placements into honors. At the same time, when the placements were judged prototypic, raters were 10% less likely to disagree.

Years of rater experience was also a significant factor, although much less so than course placement or prototypicality. Because the pool of test raters was so small, it is likely that the findings here are an artifact of particular rater pairs and less truly linked to experience, but generally speaking, the rater pairs least likely to disagree with each other were those with the greatest experience. A rater pair where each rater had been scoring tests for more than three years (high experience) was 8% less likely to disagree with each other. The fact that Rater 1's years of experience exerted a stronger statistical influence than Rater 2's suggests that these findings must be treated with extreme caution, since in practice, with independent rating, there is no reason for the order of rating to affect outcomes.

Factors Affecting Placement Decisions

This study offers some intriguing evidence about the factors that influence placement rating. Overall, the results suggest that the Washington State assumption that prototypical essays are easily identified by appropriately

trained raters is borne out. Overall, for all courses, raters were less likely to disagree over placements judged prototypic, which supports Wyche-Smith and Haswell's assumptions.

Influence of the Course in Question

In particular, the consistently high rates of agreement, regardless of meeting frequency, about placements into first year composition suggest that that course has special status in the placement system. The fact that more than half—in many semesters more than 60%—of entering students place into first year composition means that first-year composition has a much broader range of students than any other course. Consequently, it is not surprising that raters, regardless of whether or not they are currently teaching first-year composition, can recognize prototypical first year composition students. To use Smith's spatial metaphor, the distance between the upper and lower course boundaries are much greater for first-year composition than any other course, and it makes sense that the large middle ground would be relatively easy to spot. And the fact that raters could spot this middle ground even if they had most recently taught basic writing or a second-semester composition course also makes some sense. Teachers who work with students in courses that are designed to lead into or immediately follow from the first-year composition course consistently work with the expectations for the two courses; in a well-articulated program with good faculty development opportunities, all teachers should have some sense of the very center.

Influence of Teaching Experience

Does teaching experience matter? While the multivariate analysis presented in Table 4 suggests that course most recently taught did not affect placement decisions, other evidence suggests that teaching experience does matter a great deal. Smith's study, in fact, found that courses taught explained all teacher disagreement, a finding not borne out here. But our experience with honors course placement vividly illustrates the ways in which teaching experience matters. Local context, in fact, makes it virtually impossible for our raters to agree on honors placements except under very tightly controlled—and even then short-lived—conditions. In 1993 and 1994, for example, raters placed only about 60 students (out of approximately 5000) into

honors composition. In any scoring system, raters are reluctant to use ends of the scale, preferring to reserve the highest and lowest scores for tests that could not possibly be any better or any worse. No doubt this compression effect had a dampening effect on honors placements.

However, our raters' reluctance to make honors placements may also be traced to feedback from the teachers of the honors sections, who were rumored to feel that too many "artsy" students were getting placed into honors comp, students who lacked the technical skill to succeed. This feedback took on the status of something like an urban legend in the department: the honors teachers were not entirely sure how the test raters were getting this impression, yet the impression itself was clearly influencing reader decisions. Worried that they were placing the wrong students into honors, the raters became reluctant to identify *any* honors placements. In a system which privileges teaching experience, it is not surprising that raters who never taught a course would have trouble identifying students who fit into it; furthermore, teachers of first-year composition may be reluctant to make honors placements, fearing that all the good writers would be pulled out of the course they do teach. In any event, since the period of study, honors placements are no longer determined by this group of test raters. Any test identified as having a chance at honors (which we now define as a test that a rater would definitely place into honors, or a test that a rater identifies as near the upper boundary of first-year composition) is sent on to honors teachers, who make the final determination. This new system seems to have eliminated reader anxiety about honors placements.

Although traditional methods of holistic scoring have been used in many circumstances to train readers with disparate experiences to rate reliably using a given scale, all efforts to train our placement raters to use the upper end of the scale were fruitless. In this particular instance in our program, local context appears to have been such a powerful force that it overrode holistic training. Even during the four years in which a holistic scoring guide was used for placement and intensive training was held regarding the upper scale points (which raters knew would likely lead to honors placements), raters could not be trained to use the high end of the scale with any frequency. I attribute this to the interaction of teaching experience with the

scoring guide (see Elbow, "Writing Assessment" 122 for a discussion of training difficulties in a portfolio context). Because the raters believed that the honors teachers did not like the placements that resulted from their use of the high end of the scale, they were reluctant to use it, and no amount of training (by me, my predecessor, and the honors teachers) affected this for any length of time. The test raters did not feel part of a teaching community that included honors, and that feeling of exclusion affected their ability to rate.

Given that the test raters do not share anything like common course boundaries for the honors course (and that some raters have difficulty imagining any test as worthy of placement into honors), it would be a monumental undertaking to train the raters to recognize and agree on honors placements. Since the course assignment system does not permit part-time faculty to teach the honors course, for the most part, it is unlikely that the group of test raters will ever have enough familiarity with the course to be able to make consistent, confident placement decisions. We responded to this situation, in the end, by creating the tier of "expert readers" for the honors course placements. The first group of raters no longer makes final placement decisions about the honors courses; rather, they create the possibility of an honors placement by referring the test to the second group of raters. In the first year of this new procedure, honors placements rose 150%, and the teachers of the honors sections report that the students in those courses are placed properly. (The first group of test raters reports relief that honors placements are no longer their sole prerogative.) This placement pattern has held steady since then.

Influence of rater training

Particularly because the regression analysis found that years of rater experience (but not meeting frequency) had a significant effect on rater agreement, the role of group cohesion and the factors that influence the creation of shared course boundaries must be the object of further study. While the daily meetings that typify holistic scoring sessions may not be necessary to maintain acceptable rates of reliability and validity, the activities that contribute to rater experience (which is also linked to teaching experience) must be analyzed. Brian Huot found that holistic rating experience conferred significant advantages on test raters, who were able to personally engage with

student essays much more effectively than raters who lacked this training. Furthermore, the raters with holistic rating experience "organized [their] past experience into a coherent set of rating strategies" ("Influence" 226), which led Huot to conclude that "it may be that holistic scoring procedures actually promote that the kind of rating process that insures a valid reading and rating of student writing...the use of a scoring rubric made it easier not only to agree with each other, but to actually score the papers" ("Influence" 228). Of course, Huot's study examined the impact of training on reading in relationship to a scoring rubric, and his findings may not necessarily generalize to reading in relationship to direct placement decisions. What sorts of experiences confer training in a placement rating setting remain to be determined.

Nonetheless, it seems clear that a sense of community does have a significant influence on rater behavior. Research by Pula and Huot explore the varying layers of community that affect raters. The raters quoted in Huot ("Influence"), Pula and Huot, and Smith echo the raters on my own campus in valuing the training and meetings associated with placement testing. To outsiders, the reading of placement tests can seem the most tedious of processes, but to test raters, it is an exciting process that involves generative and wide-ranging conversations about expectations, curriculum, and student performance. However, the fact that meeting frequency had no impact on placement decisions for first-year composition, and affected basic writing and honors placement decisions in opposite ways, suggests that this variable requires further research. Training can occur in a variety of ways, and it is possible that new models of training can emerge from these expert scoring systems.

We need to determine what factors contribute to raters' ability to rate in an expert scoring system. Smith's study found that test raters from other universities were able, when provided with information about Pittsburgh's courses, to achieve respectable levels (72%) of agreement with Pitt's own raters (171). He attributes this finding to the effect of a writing instructor discourse community (studied in greater depth by Pula and Huot). A similar effect may explain the success of Portnet, an electronic discussion forum that brings together teachers from varied programs to discuss student portfolios. Michael Allen's report of Portnet's first year noted that the participants

found that they "had more agreement than disagreement" on scores for the portfolios discussed on line (80); he offers a range of explanations for this, ranging from the self-selection of participants with shared values, to the expertise of the group (all experienced teacher/administrators), to the nature of e-mail, to the fact of well-articulated program standards. In a co-authored essay published later in the project, various Portnet participants argue for various explanations of the agreement (Allen et al.)

In other contexts, too, wide-ranging agreements about the nature of first-year composition appears. For the past two years, a working group drawn from the Council of Writing Program Administrators has been working on an outcomes statement for first-year composition. This statement, which has been presented and developed on in forums, workshops, and panels at recent CCCC and WPA meetings, has evolved over time; teachers and administrators from a variety of schools in all parts of the US have found it an exciting experience to define a common core of outcomes for this course which is so central to our discipline. While local standards and the particulars of curriculum may differ from place to place, there is a surprising amount of agreement about this course, and this is manifest both in our local findings and in experiences such as Portnet.

On the whole, there is much potential for the exploration of how communities are maintained. Faculty development opportunities or articulate, thoughtful grading rubrics may be of as much use in generating valid placements as rating training sessions. The impact of rater meetings on placement test rating is likewise an area that will bear further study. The small number of raters who participated in this study mean that the curious impact of meeting frequency on rater agreement must be analyzed with caution. Furthermore, it must be noted that the same raters participated throughout the study, so it is possible that there was a cumulative impact of all meeting strategies. That the meeting frequency did not affect agreement for all courses in the same way only deepens the mystery about the differences between rater attitudes toward basic writing and honors. Further inquiry is needed to determine the relationships among agreement, faculty development opportunities, rater training, and meetings.

Practical Implications

Both practical and theoretical considerations have emerged from this study. As the discussion of honors placements illustrates, one practical upshot of this study was a reconceptualization of the procedures for test readings. Our experience with the honors placements and the prototypical rating suggests that one general guideline that has emerged from this research is that all tests need not be treated the same.

Before the Washington State system was introduced into national discussion, many placement test leaders assumed that all tests needed at least two readings to ensure reliability (it should be noted, however, that in recent years the Advanced Placement exam has been using only one rater for some exams). But if raters are consistently able to recognize first-year composition placements, is a second reader necessary on all those tests? The data suggest that multiple readings are not necessary for all tests. When the first rater of an exam placed the student into first year writing composition, that judgment was supported about 93% of the time (either by the second reader or by the third). Moreover, this percentage remained remarkably stable, whether the placement was regarded as prototypical or resulted from an "in-between" judgment call. In consultation with the basic writing course coordinator, who wanted to move some students out of basic writing and into first-year composition, I changed our scoring system so that tests placed into first-year composition by Rater 1 do not require further reading. This had the practical effect of increasing first-year composition placements somewhat, which fit with our program goal of giving students who seemed at the very high end of basic writing the chance to succeed in first-year composition instead.

In another innovation related to those in use at Washington State, the enlarged pool of placement raters achieved by inviting honors teachers to read some tests also offers us ways to tailor placement practices to the test in question. If prototypic placements are easier to identify (for appropriately trained raters), non-prototypic placements are harder. A system which gives some tests fewer reads than others allows the raters to devote more of their time and energy to making the difficult decisions. One important way in which not all tests need be treated the same is that not all tests demand the same of raters. Some tests are hard to rate, and our system now acknowledges that.

Is anything global about local assessment?

Clearly, this study had practical benefits for the placement system on my campus, but it also raises theoretical questions about the very nature of placement decisions. All assessments should be fit to their context, but none more so than placement testing, which determines which courses students will take. Any placement program should fit the needs of particular campuses; my campus needs to test daily, while another campus' summer testing program allows for other arrangements. But local placement experiences must be pooled in order to further test the theoretical assumptions that are the basis for entry assessments. The overlap in findings between this study, Smith's, and Haswell and Wyche-Smith's work suggests that while placement is local, some theories are global. That an open-admissions urban university, and a more selective urban and more selective rural university discover that qualified test raters can reliably identify prototypic first-year composition placements invites further study. Why is it that these placements are more easily decided? Why do courses most recently taught by raters seem to affect decisions about some placements on my campus, but not all? How do faculty form their notions of course boundaries and prototypic texts? Further research into these questions will enable placement testing to move into a more central position in the literature on writing assessment.

These questions also invite scholars to take the growing literature on validity and examine it in light of placement testing. Portfolio scholars have called on writing teachers to engage in assessment practices which examine and enact the values of the classroom (Elbow, "Foreword" and "Writing Assessment"); entry placement assessments as well must be considered in light of the values of the writing program. Portfolio assessment's popularity can be traced in part to the close connection between teaching and assessing; we must seek to create a similar link between teaching and placement testing. In particular, recent developments in assessment scholarship invite us to assess our own assessments. Peter Elbow, long an advocate for finding ways to increase the authenticity of writing assessments by opposing what he sees as reductive scoring practices, suggests that concerns for teaching should lead to the abolition of placement tests ("Do It Better" 130-131) and that at the very least, scoring procedures should be changed so that pro-

grams spend more time assessing problematic texts and less time assessing easy-to-label texts. But placement tests are defended by others who see them as a system for protecting the needs of at-risk students and offering opportunities that would otherwise be lost (see White, "Importance"). It is clear that while teacherly impulses may lead to broad agreement in some cases, there is significant disagreement among writing teachers about the best approaches to providing students with appropriate amounts of writing instruction. Placement testing practices can serve as fertile ground for research into these conflicts, for no other type of writing assessment so directly addresses the question of how we know what type of writing instruction we should provide for our students.

Controversies among writing teachers about the value of particular assessments (such as placement tests) are rooted in differing assumptions about the purposes and results of assessment. White outlines conflicting agendas of writing teachers, researchers and theorists, and testing firms and governing bodies, and students, arguing that the best future for writing assessment lies in "negotiating and compromising among the interest groups involved" ("Power" 24; see also "Writing Assessment"). In particular, we need to develop more complex ways to approach notions of validity and reliability in placement testing, to better understand the ways test results are formed in a community of raters and to better ensure that our assessment efforts will be meaningful to the varied audiences (such as campus administrators or state legislators) who will take an interest in them.

Of particular importance in this effort is the emerging literature on validity and performance assessment. Lee Cronbach suggests that validity is best understood not as an inherent property of a test, but as an argument to be made about the use of a test: "the [validity] argument must link concepts, evidence, social and personal consequences, and values" (4). Cronbach's call for the examination of evidence about the concepts tested, the evidence available, and classroom and program values echoes the scholarship on portfolios in the classroom. The rise of a constructivist assessment paradigm challenges us to examine the relationship between theory and practice and to see it as something in constant dialectic (Guba 26); Pamela Moss argues that the examination of varying assessment paradigms in light of each other

will produce critical reflection necessary to better theory and practice. The implications for placement tests are numerous. Since Smith's work offers one of a very few published examples of an attempt to validate placement scores, there is ample room for further work. Further study is required to describe the dialectic relationship between teaching experience and placement decisions, to understand the ways in which the validity of placement tests can be understood and argued, and to determine the principles that can be shared across diverse campuses. The data presented here, in conjunction with the WSU and Pittsburgh data, do not permit easy generalizations about the relationship of teaching experience to test rating.

Maurice Scharton's recent essay "The Politics of Validity" lays out relationships between varying professional belief systems and approaches to validity. In particular, he identifies an "instructional perspective" on validity which seeks "to change the political situation in assessment so that the classroom teacher rather than the institution wields the real power" (56). While Scharton identifies the instructional perspective with the portfolio movement, his formulation can be extended to include the developments in placement testing that seek to directly include teacherly expertise in the assignment of placements. In order to reconcile the instructional perspective with "programmatic perspective" (aligned with researchers and psychometricians), Scharton suggests that discussions of validity be rooted in questions such as "what cooperative measures have assessment and curricular designers undertaken to ensure integration between assessment and curricular content?" (75). A question that might be usefully added to Scharton's list might be "what measures have assessment and curricular designers undertaken to ensure that those producing the scores feel that all their expertise comes to bear on the scores?"

But the introduction of teacherly expertise into placement testing raises important issues of reliability. As Edward White noted in a review of this essay, "reliability matters." An unreliable assessment cannot be said to serve students needs (although Peter Elbow argues that it is legitimate to sacrifice reliability for validity; see "Foreword"). If, by designing placement systems that privilege teachers' classroom experience, administrators are setting up systems in which raters evaluations of essays are all over the map, unrelated

to clearly defined placement goals, students' needs are not served no matter how well the raters feel the system acknowledges their expertise. But the data reported here, and in Smith's study of the Pittsburgh program, suggest that such fears are not well-founded. But reliability does matter, and it bears some further discussion. Most importantly, we must acknowledge that the very notion of reliability has not been defined or operationalized consistently in the field. An important task for future research will be to develop approaches to reliability that allow for better comparisons between systems.

The variability in reports of reliability has been well-documented by Roger Cherry and Paul Meyer and more recently by Doug Shale. In brief, these researchers identify vague definitions of reliability (often confused with validity), over-emphasis on inter-rater reliability, and lack of agreement on statistical methods for computing reliability as some problems in the field. Shale extends his critique to include an unfailing reliance on classical test theory and proposes instead a new theoretical formulation based on generalizability theory. Shale argues that that generalizability theory better permits us to acknowledge the fact that essay raters will always vary in their responses to some texts (93) and to seek "acceptable levels of consistency for assessments of writing" (94).

The particular question facing any assessment leader is "how much disagreement is too much disagreement"? White notes that, in a good system, holistic scoring methods lead to the scoring of many tests "on a six-point scale with about 95% agreement on scores within one point" ("Apologia" 40). In comparison, the data reported here, with overall agreement rates of 82% on placements (with agreement rates for particular course placements ranging from 56% to 93%) seem low. Yet Smith reports that in his study, agreement rates of 72% met "the minimum acceptable level of agreement" when he compared judgements of different rater sets (171). What accounts for these discrepancies? Differences in scale points between a 6 point holistic scale, a four-course system at Pitt, and a 3 course system at IUPUI? And how are individual administrators to establish that minimum level of agreement? Smith argues that rater disagreement does not lead to lack of reliability (173) and that rater disagreement rather points to the fact that placement testing of necessity seeks to put students into particular cat-

egories (courses) that may not exactly match students' writing abilities; students writing abilities do not necessarily develop in ways demarcated by curriculum in a given school, and thus some students may not neatly fit into any available composition course. Shales' call for reconsidering reliability invites ways to value natural and normal tendencies for rater judgements to vary in clearly defined discourse communities (such as the communities of teachers studied here and at Pittsburgh). The administrative and scholarly challenge is to learn more about how such communities are defined and maintained, explicitly and implicitly.

Conclusion

This study represents the sort of local inquiry that builds knowledge in two ways. First, it builds local knowledge by allowing test administrators to get a better picture of the factors which affect placement decisions in their program and demonstrates the use of empirical research in program reform. Second, it offers a small piece of a puzzle that needs to be assembled by pooling local data from various sources. Placement testing ultimately needs to be looked at as simultaneously local and national; tensions between the local findings and theoretical positions need to be negotiated and evaluated. Further study, both within my program and across campuses, will help determine the extent to which the local findings I report here are anomalous.

Placement testing has lingered for too long in the shadows of classroom assessments. At a time when new forms of placement tests are being devised to meet administrative challenges, and when writing assessment theory is developing provocative ways of viewing test validity, the time is ripe for new studies of the effectiveness of placement testing and the forces that lead to sound placement decisions.

Notes

1. Dan Royer and Roger Gillies of Grand Valley State University have developed yet another model, directed self-placement, described in "Directed Self-Placement: An Attitude of Orientation." Since directed self-placement obviates the need for scoring of placement tests, I do not consider it here.

2. This research was funded by a grant from the Council of Writing Program Administrators. Thanks are due to Ellen Ann Andersen and Kathleen Blake Yancey for their generous readings of earlier drafts of this work.
3. William Smith is no longer at the University of Pittsburgh, but the testing model he introduced is still in use (with some modifications). The Washington State system is in use as described here, although Richard Haswell and Susan Wyche-Smith have left that university.
4. Because the dependent variable in this analysis is dichotomous (0 = rater agreement, 1 = rater disagreement), logistic regression is preferable to ordinary least squares (OLS) regression. What OLS lacks in statistical precision, however, it makes up for in ease of presentation. Therefore, I discuss the results of an OLS regression in the text, but include the logistic regression in Appendix A.

Appendix A

Logistic Regression of Rater Disagreement

	Parameter Estimate	Standard Error	p-value
<i>Course Placement</i>			
Basic Writing	.939	(.133)	.000
Honors	1.29	(.201)	.000
Prototypic	-.65	(.127)	.000
<i>Meeting Frequency</i>			
Daily	-.175	(.115)	.130
Weekly	.153	(.144)	.287
<i>Course Most Recently Taught</i>			
First Year Composition	.042	(.141)	.768
Basic Writing	-.242	(.218)	.267
<i>Rating Experience</i>			
Rater 1	-.267	(.066)	.000
Rater 2	-.128	(.065)	.048
Constant	-.935	(.203)	.000
-2 Log Likelihood	2544.7		
Model Chi-Square	133.0		
% Correctly predicted	81.9		
Number of cases	2842		

Works Cited

- Allen, Michael S. "Valuing Differences: Portnet's First Year." *Assessing Writing* 2.1(1995): 67-90.
- , Bill Condon, Marcia Dickson, Cheryl Forbes, George Meese, and Kathleen Blake Yancey. "Portfolios, WAC, E-mail, and Assessment: An Inquiry on Portnet." *Situating Portfolios: Four Perspectives*. Ed. Kathleen Blake Yancey and Irwin Weiser. Logan: Utah State UP, 1997. 370-384.
- Barritt, Loren, Patricia Lambert Stock, and Francelia Clark. "Researching Practice: Evaluating Assessment Essays." *College Composition and Communication* 38 (1986): 67-90.
- Broad, Bob. "Reciprocal Authorities in Communal Writing Assessment: Constructing Textual Value within a 'New Politics of Inquiry.'" *Assessing Writing* 4 (1997): 133-167.
- Cherry, Roger, and Paul Meyer. "Reliability Issues in Holistic Assessment." *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations*. Ed. Michael M. Williamson and Brian Huot. Cresskill, NJ: Hampton, 1993. 109-142.
- Cooper, Charles. "Holistic Evaluation of Writing." *Evaluating Writing: Describing, Measuring, Judging*. Ed. Charles Cooper and Lee Odell. Urbana: NCTE, 1977. 3-31.
- Cronbach, Lee. "Five Perspectives on Validity Argument." *Test Validity*. Ed. Howard Wainer and Henry Braun. Hillsdale, NJ: Erlbaum, 1988. 3-17.
- Elbow, Peter. Foreword. *Portfolios: Process and Product*. Ed. Marcia Dickson and Pat Belanoff. Portsmouth: Heinemann, 1991.
- . "Writing Assessment: Do It Better; Do It Less." *Assessment of Writing: Politics, Policies, Practices*. Ed. Edward M. White, William D. Lutz, and Sandra Kamusikiri. New York: MLA, 1996. 120-134.
- . "Writing Assessment in the Twenty-First Century." *Composition in the Twenty-First Century: Crisis and Change*. Ed. Lynn Z. Bloom, D.A. Daiker, and Edward M. White. Carbondale: SIUP, 1996. 83-100.
- Guba, Egon. "The Alternative Paradigm Dialog." *The Paradigm Dialog*. Ed. E. Guba. Newbury Park, CA: Sage, 1990. 17-27.
- Haswell, Richard, and Susan Wyche-Smith. "A Two-Tiered Rating Procedure for Placement Essays." *Assessment in Practice: Putting Principles to Work on College Campuses*. Ed. Trudy Banta. San Francisco: Jossey-Bass, 1995. 204-7.
- Huot, Brian. "The Influence of Holistic Scoring Procedures on Reading and Rating Student Essays." *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations*. Ed. Michael M. Williamson and Brian Huot. Cresskill, NJ: Hampton, 1993. 206-237.
- . "The Literature of Direct Writing Assessment: Major Concerns and Prevailing Trends." *Review of Educational Research* 60 (1990): 237-263.
- . "A Survey of College and University Writing Placement Practices." *Writing Program Administration* 17.3 (1994): 49-65.
- . "Toward a New Theory of Writing Assessment." *College Composition and Communication* 47(1996): 549-566.
- Lloyd-Jones, Richard. "Primary Trait Scoring." *Evaluating Writing: Describing, Measuring, Judging*. Ed. Charles Cooper and Lee Odell. Urbana: NCTE, 1977. 33-68
- Moss, Pamela. "Enlarging the Dialogue in Educational Measurement: Voices from Interpretive Research Traditions." *Educational Researcher* 26.1(1996): 20-28.
- . "Shifting Conceptions of Validity in Educational Measurement: Implications for Performance Assessment." *Review of Educational Research* 62.3(1992): 229-258.
- Murphy, Sandra, et al. *Report to the CCCC Executive Committee: Survey of Postsecondary Writing Assessment Practices*. 1993
- Pula, Judith and Brian Huot. "A Model of Background Influences on Holistic Raters." *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations*. Ed. Michael M. Williamson and Brian Huot. Cresskill, NJ: Hampton, 1993. 237-265.
- Royer, Daniel J. and Roger Gilles. "Directed Self-Placement: An Attitude of Orientation." *College Composition and Communication* September, 50.1 (1998): 54-70.
- Scharton, Maurice. "The Politics of Validity." *Assessment of Writing: Politics, Policies, Practices*. Ed. Edward M. White, William D. Lutz, and Sandra Kamusikiri. New York: MLA, 1996. 53-75.

- Shale, Doug. "Essay Reliability: Form and Meaning." *Assessment of Writing: Politics, Policies, Practices*. Ed. Edward M. White, William D. Lutz, and Sandra Kamusikiri. New York: MLA, 1996. 76-96.
- Smith, William. "Assessing the Reliability and Adequacy of Using Holistic Scoring of Essays as a College Composition Placement Technique." *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations*. Ed. Michael M. Williamson and Brian Huot. Cresskill, NJ: Hampton, 1993. 142-205.
- White, Edward. "Apologia for the Timed Impromptu Essay Test." *College Composition and Communication* 46 (1995): 30-45.
- . "The Importance of Placement and Basic Studies: Helping Students Succeed Under the New Elitism." *Journal of Basic Writing* 14.2 (1995): 75-84.
- . "Holistic Scoring: Past Triumphs, Future Challenges." *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations*. Ed. Michael M. Williamson and Brian Huot. Cresskill, NJ: Hampton, 1993. 79-108.
- . "Power and Agenda Setting in Writing Assessment." *Assessment of Writing: Politics, Policies, Practices*. Ed. Edward M. White, William D. Lutz, and Sandra Kamusikiri. New York: LA, 1996. 9-24.
- . *Teaching and Assessing Writing*. 2nd ed. San Francisco: Jossey-Bass, 1995.
- . "Writing Assessment Beyond the Classroom: Will Writing Teachers Play a Role?" *Composition in the Twenty-First Century: Crisis and Change*. Ed. Lynn Z. Bloom, D.A. Daiker, and Edward M. White. Carbondale: SIUP, 1996. 101-115.