Journal of Writing Assessment Vol. 1, No. 1, pp. 47-65

Moving Beyond Holistic Scoring Through Validity Inquiry

PEGGY O'NEILL
Loyola College[1]

This article re-examines the research into placement that William L. Smith did at the University of Pittsburgh during the 1980s and 1990s by situating Smith's work within the larger context of educational measurement theories, placement testing, and holistic scoring. I present the series of research studies that Smith conducted into Pitt's placement test as a case study in validation inquiry, arguing that his approach serves as a model for those who direct writing assessments. The implications of Smith's research reach beyond placement into first year composition: by approaching local writing assessment needs as Smith did, writing assessment professionals not only can create more effective assessments, but they also can contribute significantly to assessment theory.

Since 1874 when Harvard introduced English composition as a subject in the battery of entrance exams prospective students completed in the application and admission process, writing assessments have become standard features of college entrance exams, playing a role in students' college curricula choices. Early writing tests—essays about literature—were typically evaluated by professors;

Peggy O'Neill, an assistant professor of writing, directs the first-year composition program at Loyola College in Maryland where she teaches a variety of undergraduate writing courses. Her scholarship focuses on writing assessment theory and practice, preparation of writing teachers, and writing program administration. She is also interested in the politics of writing programs and the disciplinarity of rhetoric and composition. Besides authoring or coauthoring several journal articles and book chapters, she has coedited *Field of Dreams: Independent Writing Programs and the Future of Composition Studies* (2002) and *Practice in Context: Situating the Work of Writing Teachers* (2002).

Direct all correspondence to: Peggy O'Neill, Department of Communication, Loyola College, Baltimore, MD 21210-2699

however, during the 20th century, this practice gave way to more scientific methods, so that educational measurement theories and practices dominated writing assessment (White, 1998; Williamson, 1993). Tests of grammar, usage, and mechanics that required little or no writing (e.g., fill-in-the-blank, multiple-choice, editing) were popular, but by mid-century, instructors of writing and directors of writing programs had become increasingly disgruntled with these exams. Impromptu essay exams re-emerged as a popular method of placing students into the first-year composition program; however, these "new" essay exams depended on holistic scoring, a "scientific" and "objective" type of evaluation of student writing.

Holistic scoring of timed essays quickly spread until it was assumed to be the best practice for placing students into the first-year writing curriculum. Edward White (1995) explained that the popularity of the impromptu essay test, which was defended by English faculty because it replaced the use of multiple-choice exams, rested on the holistic scoring procedures, which were cost-efficient and produced valid and reliable results. Writing assessment practitioners and scholars assumed essay testing results valid because the test demanded students write instead of filling in scantron answer sheets, and reliability rates were acceptable as long as the readings were properly managed (White, 1998). The acceptance of holistic scoring as the cornerstone for direct writing assessment (Wolcott, 1998) continued as compositionists began to experiment—and favor—portfolios over impromptu essays for writing assessment. For example, Miami University used holistic scoring to evaluate portfolios submitted for advance placement in first-year composition (Beck, Dautermann, Miller, Murray, & Powell, 1997). Holistic scoring of essays or portfolios typically went unquestioned as long as the interrater reliability coefficients were acceptable. Smith's (1992) response to the essay placement exam he inherited on his arrival at the University of Pittsburgh typifies this position: "It seemed to work, so there was no impetus to examine it, let alone change it. The incoming students were placed into our courses efficiently and with what appeared to be tolerable numbers of errors" (p. 314). However, although most compositionists remained complacent about using holistic scoring of timed essays for placement testing, Smith began to question the practice at his institution. Smith's tinkering with the placement testing during his association with the University of Pittsburgh's writing program produced not only several published research reports and numerous conference presentations but also demonstrated how systematic, ongoing validity research functions to enhance a particular local test and contributes—both theoretically and practically—to the scholarship of writing assessment.

By situating Smith's work within the larger context of educational measurement theories, placement testing, and holistic scoring and presenting it as a case study of validity inquiry, I argue that by approaching local assessment needs as Smith did, compositionists can create better assessments while contributing significantly to writing assessment theory.

Validity

According to Smith (1998), there is a "paucity of validation research" (p. 3) in writing assessment, which stems from several different but interrelated problems: a lack of understanding of key concepts such as validity and reliability; an overemphasis on achieving reliability; a lack of understanding of what validation inquiry entails; and a failure to articulate the theoretical constructs underlying writing assessments. Correcting these deficiencies in the composition literature on assessment has begun (Huot, 1996; Moss, 1994, 1998; Smith, 1992, 1993, 1998; Williamson, 1993) but the confusion still exists, especially in our understanding of validity.

Validity has been—and continues to be—misconstrued in most of composition's assessment literature. White (1995), one of the most prolific voices in composition's assessment community, wrote, "Validity means honesty: the assessment is demonstrably measuring what it claims to measure." (p. 40). In the revised edition of his popular book, *Teaching and Assessing Writing,* White (1998) stated: "Although validity is a complex issue—colleges offer advanced courses in it—one simple concept lies behind the complexity: honesty. Validity in measurement means that you are measuring what you say you are measuring, not something else, and that you have really thought through the importance of your measurement in considerable detail" (p. 10). Even more recent discussions have continued this misperception despite measurement theory and writing assessment literature that contradicts these simplified definitions. For example, Harrington (1998) explained validity this way: "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)" (p. 59). Yancey (1999) asserted: "Validity means that you measure what you intend to measure" (p. 487). And Shane Borrowman (1999) quoted White when defining validity: "According to Edward M. White, 'Validity . . . has to do with honesty and accuracy, with a demonstrated connection between what a test proclaims it is measuring and what it in fact measures'" (p. 9). In discussing their self-placement system, Royer and Gilles (1998) sidestepped the issue of validity for the most part: "Our old concerns about validity and reliability are now replaced with something akin to 'rightness'" (p. 62). Although Royer and Gilles, Harrington (1998), and others acknowledged more complex considerations of validity, in addressing validity issues of their specific placement systems, they resorted to simplistic and faulty conceptions. Validity, however, is a complex notion in assessment that should not be distorted or simplified to fit individual agendas, nor should it be reduced to a one-sentence sound bite.

In the educational measurement community, debates and discussions about validity have been ongoing. Two of the most influential voices in these discussions have been Cronbach and Messick (Moss, 1992, 1995; Shephard 1993), who each has written about the complexity of validity's theoretical nuances and practical applications for over three decades. Although there has been considerable debate in the assessment community, several scholars such as Lorrie Shepard (1997) and Pamela Moss (1992), argue that the *Standards for Educational and Psychological Testing,* which are the standards for research and measurement endorsed by the American

Psychological Association, the American Educational Research Association, and the National Council on Measurements Used in Education, and the scholarly literature do in fact support the unified, complex notions of validity that have evolved more recently. According to Cronbach (1988), who was instrumental in drafting the original *Standards*, validity "must link concepts, evidence, social and personal consequences and values" (p. 4). Messick (1989) argued that validity uses "integrated evaluative judgment," supported by empirical evidence and theoretical rationales, "to support the adequacy and appropriateness of inferences and actions based on test scores and modes of assessment" (p. 5). In other words, validation arguments are rhetorical constructs that draw from all the available means of support. Validation studies include issues of reliability, construct definitions, consequences, and other empirical and sociopolitical evidence. Huot (1996), who drew on the work of Cronbach and Messick, concluded that "in writing assessment, the validity of the test must include a recognizable and supportable theoretical foundation as well as empirical data of students' work" (p. 550). Valid writing assessments, he continued, "need input from the scholarly literature about the teaching and learning of writing" (p. 550). In validating a writing assessment, Huot recommended that writing researchers also include inquiry into the use of the assessment results. These conceptions of validity, argued Huot, "look beyond the assessment measures themselves and demand that a valid procedure for assessing writing must have positive impact and consequences for the teaching and learning of writing" (p. 551). In placement testing, validation demands determining the adequacy of placement as well as investigation into other aspects of the test, such as the testing and scoring procedures, to determine if students are being placed in the course which best fits their needs. Ensuring adequate placement should allow more effective teaching and learning because teachers will be able to better meet the needs of students.

Although validity theory is the overarching issue in assessment, reliability has most often dominated discussions of writing assessment, especially in terms of holistic scoring. As with validity, misperceptions about reliability have a long history in direct writing assessment. Reliability has been construed as a simplistic notion in most of the holistic scoring literature, which has been marked by an inconsistency and confusion in defining and calculating reliability (see Cherry & Meyer for a detailed discussion). In many cases, reliability has been reduced to interrater reliability, the agreement between two independent readers, although "interrater reliability alone cannot establish holistic assessment as a reliable or valid procedure" (Cherry & Meyer, 1993, p. 114). In fact, many different facets of reliability are at issue in rating essays such as intrarater reliability, the degree to which raters agree with themselves; rater set reliability, the consistency of rating of two primary readers that constitute a set; and instrument reliability, the consistency of the test itself across successive administrations, which takes into account students, tests and scoring as potential sources of error (Cherry & Meyer, 1993). As Cherry and Meyer explained, "Regardless of how consistently raters assign scores to written texts, if the writing prompt (the test) is faulty or if examinees do not respond consistently to it, the holistic scores will not reliably reflect writing ability" (p. 115).

Coupled with the ongoing misunderstanding about what reliability entails is a failure to acknowledge that reliability contributes to a validity argument but is not itself enough to validate the results of a test. In fact, Moss (1994) turned to interpretive research traditions such as hermeneutics to argue for the inappropriateness of reliability as a key part of validation in some types of assessment. According to Moss, traditional assessment privileges standardization but it is inadequate in evaluating complex performances such as reading and writing. A hermeneutic approach would include "holistic, integrative interpretations" that would "privilege readers who are most knowledgeable about the context in which the assessment occurs," and "ground those interpretations not only in textual and contextual evidence available, but also in a rational debate among the community of interpreters" (p. 7). This approach to writing assessment would support the processes and theories associated with literacy, leading to more theoretical alignment between actual literate practice and the assessment of it. Moreover, a hermeneutic approach undermines the quest for "objective" rating of essays that supports the proliferation of holistic scoring as the preferred procedure for direct writing assessment.

Besides—and maybe because of—these problems with key assessment concepts of reliability and validity, there is a lack of rigorous composition research into placement methods. Although there are not many models to follow, Moss (1998) explained that in composition placement

> a sound program of validity research begins with a clear statement of both the purpose and the intended interpretation or meaning of test scores and then examines, through logical analysis, the coherence of tests with that understanding. Without a clear sense of how validity and validation inquiry plays into the development and evaluation of a placement test, it is not possible to be sure that students are being placed into the appropriate course. (p. 117)

In the case of a placement exam, logical analysis of coherence must also encompass an understanding of the different courses as well as the outcome measures used to evaluate success in those courses. Directors of placement tests need to systematically collect a variety of data such as raters' decisions and interviews and surveys of participants, and analyze the data through multiple perspectives. Moss (1998) also suggested that validity inquiry should include other methods such as critical linguistics (linguistic analyses of discourse that surrounds an event) or ethnographic studies (participant-observer research). Validity research involves a dynamic process that requires an examination of procedures and results, use of this information to revise and improve assessment practices, and an examination of revised practices in a never-ending feedback loop. In short, validity inquiry should be embedded in the assessment process itself, ongoing and useful, responsive to local needs, contexts or changes, something that is never really completed.

The work that William L. Smith, along with a cadre of graduate students, did for more than a decade during his tenure as director of testing for the University of Pittsburgh composition program during the 1980s and early 1990s is an example of how systematic, ongoing validity inquiry can not only lead to better—more valid—local assessment but also contribute to the larger field of writing assessment.

Placing Students Versus Holistic Scoring

A key to understanding the validity research Smith conducted is to understand the difference between holistic scoring—a procedure for evaluating texts—and placement—the decision that is made about the writer based on the results of an evaluation. Although this distinction may seem obvious now, it wasn't always so clearly understood.

Traditionally, compositionists have talked about writing assessments in terms of direct and indirect tests. Indirect tests do not use student writing as part of the test but rather extrapolate "writing ability or potential" indirectly from, for example, the students' SAT or ACT scores or other multiple-choice tests of language use such as the computer adaptive COMPASS or ACCUPLACER. The most recent published surveys of placement (Huot, 1994; Murphy et al., 1993) demonstrate that multiple-choice tests are still very popular methods of composition placement although students do not do any actual writing. Compositionists tend to favor direct measures because they use student writing as the basis for the assessment. Samples of student writing may be collected through impromptu essays given during a testing period, online writing submitted via the Internet, or portfolios of school or self-sponsored writing. Although the sampling methods may vary, most large-scale direct assessments are evaluated through holistic scoring (see White, 1998; Wolcott, 1998).

Although one way to describe writing assessments is by the sampling method (direct or indirect), a more productive way to look at an assessment is through its purpose: Why are we assessing student writing? Possible responses include program assessment, student proficiency, or placement. Another way to see a test is through its effect: What are the consequences of this test to students, programs, teachers? By posing these sorts of questions, we move beyond the sampling method to a more productive framework for identifying similarities and differences.

Placement testing that uses writing samples has often been conflated with holistic scoring. For example, placement often uses timed impromptu essays to collect a writing sample, much like large-scale assessments such as National Assessment of Educational Progress. Additionally, placement rating is like holistic rating in that the readers use the basic holistic method: a single, quick reading leading to a single, overall judgment. Additionally, both types of assessment generally use two independent raters as the basic decision-making unit. However, placement rating is unlike holistic in some very important ways:

1. In holistic rating, the meaning of the points on the scale are internally derived; it depends on both the range finders and the range of the essays in the set to be rated. In placement rating, the points on the scale are externally derived because the scale is determined by the institutional context: the curriculum, the assumptions about composition, and the purposes of each course. The particular set of essays being rated does not influence these conditions and does not determine the scale.
2. In holistic scoring, an interval scale is used, which means that the distance between points on the scale is the same. That is, the range from

Point 1 to Point 2 is equal to the range from Point 2 to Point 3. More importantly, the difference between a 2.5 and a 3.5 is equal to the distance between a 1.5 and a 2.5. A rater holistically scoring texts is ranking and comparing the texts along the scale. Because the text is being compared to the others in the set, summing of the primary raters' score or averaging of them is acceptable. A split-resolver's score can be averaged or substituted without problem. In other words, texts can receive scores along the continuum of the scale. However, in placement a categorical—also known as an ordinal—scale is used, which means that the distances between points is often more varied; consequently, the distance between the midpoints is not equal.

The placement scale is actually determined by the curriculum with each scale point representing a curricular choice (e.g., basic writing, composition 1, honors composition). The range for the point that represents the standard first-year course is usually wider than other courses such as developmental or honors. Because the scale is categorical, texts need to be slotted into one category or another; therefore, differences among raters cannot simply be averaged because between-course scores can result. In fact, it is to be expected that some students, through their writing sample, will exhibit characteristics of more than one course, not fitting neatly into any course (or any point on the scale, any one category) although they have to be placed into one course.

3. In holistic scoring, the scale is defined by the set of student texts being evaluated; therefore, the texts "fit" the scale. In placement, however, the scale is pre-set by the curriculum, so the students have to fit the scale, which isn't always the case. This feature affects the distribution of students. Because the scale is not set by the pool being evaluated but is predetermined, the distribution of students along the scale should vary from year to year. If the distribution does remain constant it is highly likely that either students are being placed in order to fill seats in classes, not to put them in the most appropriate class, or there is a very stable pool of students.
4. In holistic scoring, the focus is on the text and locating the text on a scale. In placement, the focus is on the student and placing the student in the appropriate course. There are very real consequences in placement and raters have explained that even if a holistic scale is used, they make judgments about students not just texts (e.g., Pula & Huot, 1993; Smith, 1993).

Holistic scoring, then, is not the most appropriate method for placement although it may be useful for other situations, for example, when the results of the test are used to evaluate a program, not individual writers.[2] In this scenario, the scale points can be determined according to what the test giver desires to learn, such as whether or not the program's outcomes are being met. In this type of testing, the features of the written text can provide answers to the research questions, and the results are reported for the group with no consequence to the individual

writers. Placement, by its very nature, has consequences for individuals and it needs to be distinguished from holistic scoring.

Besides all of these differences, several issues that influence its design and implementation are unique to placement. Practical issues, such as administration and processing of placement essays (especially in reference to turn-around time), have to be negotiated with other campus parties, such as the orientation coordinator and the advisement center. Composition curricula, enrollment patterns, first-year student demographics, orientation demands, and funding may also be influential factors in designing individual placement systems. Other elements such as the pool of available readers or the size of the set to be read also need to be considered.

All placement methods, however, assume that different courses are needed to meet the needs of different students, and all acknowledge the need for some type of sorting mechanism for matching students to the appropriate courses. Because our placement methods sort students, as professionals it is imperative that we validate our placement assessments:

> [W]e have an obligation to make certain (i.e., conduct research) that our testing is fair and valid, in elicitation methodology, in the scales used, in the ways we make judgments on the writings, in the ways we analyze, interpret, and use the results, and in the ways and forms in which we publish those results . . . only through rigorous forms of validation research can we really construct assessments that accurately and ethically assess our students and programs. (Smith, 1998, p. 3)

Besides ethical and professional obligations, we should be concerned with our ability to legally defend our assessments if challenged. William Lutz (1996) explained that few academics realize that there are enough legal precedents to indicate there is liability associated with assessment, even institutional testing such as placement and exit. To be prepared for a legal challenge and to ensure we act ethically and professionally, Lutz (1996) and Smith (1998) recommended similar approaches: We need to conduct systematic, ongoing research into our methods, procedures, and programs. In most cases, however, very little rigorous research has been conducted to determine the validity of placement decisions.

A Case Study of Validation Inquiry

As Smith (1992) explained, when he started at Pitt he inherited a method for placing students that was standard and consistent with what other universities used. Most of the placement occurred over the summer during orientation sessions. Students wrote their essays in large group sessions spread over the summer months. Because the composition program was based on the interrelation of reading and writing, students were given a passage to read and a series of questions designed to focus their response. The prompts closely resembled the assignments students experienced in the composition courses. Students were given 2 hours to complete the task. The essays were rated immediately after students finished by composition faculty who were trained raters and experienced teachers.

The rating system consisted of two primary raters who scored the essay independently. The scale corresponded to the curriculum: a rating of A, B, C, coincided with the three composition courses, and D rating indicated "exempt."[3] If the two primary raters agreed, the student was placed into that course; if they disagreed, a third rater, a "split-resolver" was used.

Although Smith acknowledged his initial complacency with the placement test he inherited, he began to feel uneasy about it so he embarked on a series of research projects, which he conducted for more than a decade. Data sources for these projects included surveys of faculty and students; interviews with students, teachers and raters; think aloud protocols of raters; analyses of rater and rater-set decisions; grade distributions; placement distributions; and statistical analyses. Although he designed and conducted a series of distinct studies, Smith found that his interpretations and conclusions depended on the accumulated knowledge and experience he garnered from the ongoing nature of his work.[4] In other words, he did not keep reworking and revising his research until he got the placement process "right"; instead, his research helped him to form new research questions, revise his research approach or focus, and revise the placement procedures. Ultimately, it led Smith to develop new placement assessment methods, which he continued to research until he left Pitt.

Determining Adequacy of Placement

Smith realized that he had no solid evidence that the placement system was working—that students were appropriately placed. Like most composition placement systems, Pitt's seemed to be adequate because the error rate—the number of misplaced students—appeared relatively low. Determining the error rate depended on an essay written during the first week of class. Based on this essay, teachers identified students they believed misplaced, and a senior faculty member would read the essays, moving students into different courses if necessary. The first-week essay check, reasoned Smith, provided only marginal evidence about how many students were misplaced. He suspected that the error rate was seriously underestimated: teachers were reluctant to have students transferred out (which meant others may transfer in), some students were absent for the first-week essay, and students' attitude toward their placement (and the specific class and teacher) may have effected their effort on the first-week essay.

Because Smith wasn't content with the procedures for determining adequacy of placement, he spent 3 years developing methods for figuring out if students were being placed adequately or not. He concluded that adequacy of placement depended on triangulating several different data sources, none of which was sufficient by itself:

- The number of students moved to a new course during the first week.
- Student's final course grades; student's impressions—collected during and especially after the course—of the degree to which the course met their needs.
- Teachers' impressions of how well the students fit the course.
- Exit exams or posttests.

Alone, each of these measures has problems. For example, grades could be influenced by factors such as attendance and promptness that are unrelated to the appropriateness of the course. An exit exam or posttest might often depend on just one writing sample that may not adequately represent students' abilities; or as at Pitt, not all courses required exit exams. After surveying teachers and students, conducting interviews with teachers and students, and analyzing the results, Smith determined that teacher perception is the single best measure of whether students belong in the course. Smith also found that teachers' perceptions of students change considerably across the course of the semester. If gathered too early in the semester, teachers don't have enough evidence on which to base their decision; if gathered too late, teacher perception correlates very highly with the students' final grades, indicating that the students' actual performance is evaluated, not their potential. Smith concluded that teacher perception data should be collected during Weeks 3 through 5 of a 15-week semester

Causes of Errors in Placement

Besides developing procedures for determining a more accurate error rate, Smith focused on investigating the possible causes for error. He identified several potential sites of error: the writing prompts, the conditions under which the students write, the writers not writing essays that accurately represent them, raters not making good decisions, and an inadequate rating scale. He also acknowledged that as the director of testing, he was another source of error because he was the one who decided on the testing procedures, hired the raters, and evaluated the system. After specifying these potential sources of error, Smith began systematic inquiry into each one.

The first area for investigation was the writing prompts. The placement prompt required students to read a short text and then respond to it. Smith and his coresearchers conducted a series of studies where they varied the format of the prompt and analyzed the writing that students produced.[5] They examined three different types of prompts with writers at different ability levels. The writers responded in class within the time period allotted for the placement. After analyzing the results for types of errors, frequency of errors, and fluency, Smith concluded that the prompt they were using was adequate because it differentiated writers appropriately for the courses offered in Pitt's writing program.

The second series of studies investigated the placement exam conditions. Did it make a significant difference if students wrote their essay in large groups or small groups? Was the 2-hour time limit a factor? Did "warm-up" exercises make a difference in students' writing? Did it matter if a "real" composition teacher, who explained Pitt's composition program to the test takers, administered the test? According to Smith's studies, time and group size were not factors while warmups and teachers made only slight differences. These differences were only apparent for the weaker writers, and the differences were not consistently positive. Smith concluded that the testing conditions were not very influential factors.

Smith next turned to looking at the writers and found that there were significant factors that influenced their performance, but that he could not control for them.

For example, Smith found (to no one's surprise) that many students—especially males—are distracted when they take the placement exam by the new setting, new people, and new freedom they encounter in their trip to campus. However, administering the test during another time was not feasible.

The rating scale was another factor that Smith had no direct control over. Placement rating scales should be determined by the courses in the composition program, so that at Pitt, the 4-point scale corresponded to the three composition courses and an exemption from composition option.

Finally, Smith's research led him to focus on the raters, which yielded not only a wealth of information but helped Smith revise the placement system.

Focusing on the Raters

Smith set out on a series of studies focusing on the raters and rating system he used for placement. The raters were teachers in the composition program who, as paid volunteers, scored placement essays during the summer. All raters had experience teaching Course C, whereas some had experience with Course A and/or B, but during a given semester teachers only taught one course (because most raters and teachers were graduate students, this was done so they had only one course preparation per term). All placement essays were read by two raters and if they disagreed, another rater (split-resolver) decided the rating. The second rater did not know the first rater's score; the split-resolver knew he or she was a split-resolver, not a primary rater. In this system, each rater was responsible for making multiple decisions: Does the student belong in Course A, Course B, Course C, or should the student be exempt from composition altogether?

In conducting his studies of raters, Smith relied for the most part on records of rater decisions collected as they rated, rater profiles of teaching experience, and think-aloud protocols. His experience with the raters and the placement system were also valuable sources of information. By keeping meticulous records of the rating decisions for each essay and re-rating of certain essays, Smith was able to collect detailed information about how raters scored. He used these procedures to conduct a series of studies that examined rater reliability (interrater and intrarater), rater-set reliability, and split-resolver rating patterns. His research led him to examine how raters' profiles correlated with reliability rates and eventually how placement decisions were influenced by raters' teaching experience. (For an in-depth account of these studies see Smith, 1993). Smith's conclusions also resulted from his willingness to re-examine and rethink his approach and the data. For example, instead of merely focusing on trying to get raters to agree more consistently, Smith looked at when raters disagreed, determined if the disagreements were reliable (they were) and then tried to figure out why. In making hypotheses and testing them, Smith did not neglect to go back to what prompted the studies in the first place: Were students' placements valid? He used the procedures for adequacy of placement, most importantly teacher perception, that he had developed to see if students were indeed being appropriately placed.

Based on these studies, Smith made some conclusions about placement research and procedures that may be useful for other researchers:

- *When raters knew they were being tested, they responded differently*: In placement research, that means that dry runs, "staged" placement sessions, or other uses of holistic scoring may not be adequate representations of what raters do in "real" placement. Recirculating essays without the raters knowledge is necessary to get an accurate sense of rater reliability.
- *Raters who are split-resolvers rate differently than when they are primary raters*: Placement for students who fall between courses is not the same as those who fit the scale more easily, which means interrater reliability is affected because raters and rating are not consistent.
- *Raters made decisions about students, instead of merely judging texts*: In think-aloud protocols and informal conversations about placement reading, raters often referred to their classrooms and the student writer instead of the text. Because raters are deciding what course a student should take, and not judging the text itself, raters can disagree about quality but agree on placement. However, disagreement, whether about quality or placement, is to be expected. Holistic scoring, on the other hand, actually tries to eliminate or minimize disagreement, focusing instead on consensus or agreement.
- *Some students didn't fit into any course:* It is reasonable to assume that not all students will fit neatly into one of the composition courses because the scale is predetermined by the curricula. This is a potent source for disagreement.
- *Using traditional methods for determining reliability did not accurately portray what raters actually did nor how reliable their judgments were*: Reliability in most writing assessments has been determined by interrater reliability alone, which represents how often raters agree with each other; however, this statistic masks other important aspects of reliability: Is a rater consistent with him or herself? Is a rater-set consistent? Are raters consistent in their disagreements? Do split-resolvers rate consistently? Unpacking reliability complicates determining whether a placement test is reliable, but it provides more information for determining if the test results are valid because it provides multiple perspectives and data, allowing the researcher to get a more nuanced understanding of what the raters and the rating process.
- *Raters' teaching experience affected their rating, perhaps even more than calibration*: Raters were all experienced teachers and depended on that experience and knowledge in determining placement. It proved to be more powerful than calibration or practice sessions in their decision making.
- *The course the rater most recently taught affected the rater's decision:* Ultimately, when comparing the rater's most recently taught course experience to their rating decisions, raters were most consistent in placing students into the course they had most recently taught. Their consistency decreased the further away they were in terms of experience from the course. For example, a teacher who most recently taught

Course C placed students into Course C more reliably than in Course B, but Course B placements were more reliable than those for Course A.

The Expert Model

Based on these conclusions, Smith changed the placement procedures to what he called the "expert model." In this system, raters were assigned to rate for one course only, the one they most recently taught. They made only a binary decision: Accept the student for their course or reject him or her. Depending on the course for which they were rating, they could reject high or reject low. The basic process was as follows:

If the first reader accepts, the next reader has the same course-taught expertise (CTE).
If a CTE-(Course) A rejects high, the next reader is CTE-C (because most students ended up in C).
If CTE-B rejects low, the next reader is a CTE-A.
If CTE-B rejects high, the next reader is CTE-C.
If CTE-C rejects low, the next reader is CTE-B.
If CTE-C rejects high, the next reader is CTE-D.

Of course, because Course D represents exemption, there can be no CTE; instead a panel of expert teachers read the essay and decided if the student should be exempted from all composition or take Course C. Inevitably, as Smith found out, some students did not fit neatly into a particular course; he called them "'tweeners" because they fell between courses. All essays were read at least twice until they were located on the following scale:

COURSE A
BETWEEN COURSES A & C
COURSE B
BETWEEN COURSES B & C
COURSE C
BETWEEN COURSE C & D (EXEMPT)
EXEMPT

Smith's research indicated that raters reliably rated tweeners between courses. Smith also found that in the traditional placement system, which used split-resolvers, tweeners' placement was affected by the split-resolvers' most recent course taught experience so that tweeners were not reliably placed. In the expert model, Smith determined that all tweeners would go to the next highest course except for those between Course C and exemption; they would take Course C. Analysis of the adequacy of placement of tweeners found that they did not have a higher failure rate than students placed directly into the course although teachers continued to identify them as marginal, not an exact fit for the class. (Interestingly,

the perception as a tweener continued once the student passed through Courses A or B and into C). The overall rate of error—the number of misplaced students—was less than 3% with the expert model, but even more importantly, the number of prototypic students for each course increased (there were less marginal students in each course).

Smith, of course, acknowledged the need for more research to test the expert model. For example, would the practice of moving tweeners to the higher course ultimately affect the teachers' perception of the prototypic student? Would the reliable placement of students through the expert model prove itself through multiple years of inquiry? Before he could address these questions, Smith left Pitt. However, his work has made a considerable contribution to not just placement research and procedures but also to writing assessment in general.

Conclusions and Implications: From Local Applications to Assessment Theory

Smith's placement research was grounded in Pitt's composition program, not necessarily universally applicable. For example, Pitt had a composition program with clearly articulated assumptions about writing and teaching writing that were shared by the faculty. Furthermore, the expert model depends on having teachers teach all sections of the same course in a semester. In many composition programs, this isn't possible so teachers may be teaching two or more of the first-year composition courses would have more than one most recently taught course, which may be a factor in their placement decisions. Although the particulars of Smith's research, conclusions, and revised placement procedures will not fit another program exactly. His conclusions and procedures can help other placement directors design studies and procedures, and the implications of the work reach beyond placement to other forms of writing assessment.

One of the most important aspects of an assessment is validity, yet it is also an area that is under researched and misunderstood in composition's assessment literature. Smith's work not only illustrates how to conduct validation research but also how writing specialists need to understand the complexities—both theoretical and practical—that validity involves. Validity inquiry needs to focus on the purpose and use of the test's results and requires more than a quantitative analysis of the results. As Moss (1994) argued, traditional standardized, objective approaches to assessment are inadequate for evaluating complex performances such as reading and writing. A hermeneutic approach would include "holistic, integrative interpretations" that would "privilege readers who are most knowledgeable about the context in which the assessment occurs," and "ground those interpretations not only in textual and contextual evidence available, but also in a rational debate among the community of interpreters" (p. 7). Smith's expert model enacted this approach: He allowed experienced, expert teachers to make holistic, integrated judgments about student placement, and he grounded these decisions with a variety of evidence and rational debate. This approach to writing assessment endorses the approach to reading and writing supported by composition scholarship, and it undermines the quest for an "objective" rating of essays that accompanies holistic scoring, the most

popular procedure for direct writing assessment. In placement testing, validity rests on determining that the students are being adequately placed, a task that is more involved than most programs acknowledge. In exit testing or competency testing, validity inquiry will take different forms. Local context, including faculty, curricula, student populations, come into play in collecting and analyzing data and building a validation argument.

Smith's work also reminds compositionists that reliability is complex and multidimensional. Composition as a field has often relied on interrater reliability in determining reliability, but that distorts the notion of reliability. Readers' disagreements are an important source of information that needs to be unpacked. Resorting to a simplified reliability coefficient can mask important aspects of a rating system, of reliability, or of validity. By examining when readers disagreed, Smith realized that readers can reliably disagree. There may also be factors that influence reliability, which Smith discovered he could control for. In Pitt's placement program, teachers' most recent course taught experience was a significant factor in reliability of ratings; in other programs, there may be other factors such as education or background. In other types of writing assessments, such as competency testing or exit testing, reliability may be influenced by different factors specific to the test's purpose, the curriculum or other contextual variables. In short, individual writing assessments and the requisite validation inquiry that should accompany them need to be sensitive to local context.

The ongoing research conducted by Smith highlights the demands of writing assessment, which is a specialized field that requires practitioners to understand composition theory as well as assessment theory. Smith's work not only legitimizes assessment work as discipline defining and knowledge-generating but also as something that demands specialized knowledge and education. Writing assessments, after all, play an important role in identifying values and assumptions about writing, evaluation, and teaching of writing. Unfortunately placement (or other assessment demands) are most often viewed as part of administration or service, requiring no specialized knowledge or education. Huot (1994) found that only 14% of schools' using direct writing assessment for placement had a director with a terminal degree in composition or publications in writing assessment. In other words, many of the professionals designing, implementing, and evaluating placement tests are not writing specialists, let alone writing assessment specialists. By allowing assessment to be controlled by professionals without the necessary knowledge and experience, we are in effect allowing our field to be dominated and defined by those outside the field.

Likewise, as long as compositionists continue to separate themselves from the larger educational assessment community (Huot, 2002), we run the risk of merely adopting assessment methods and approaches that are inconsistent with our assumptions that literacy is a complex, contextual activity. Writing assessment specialists need to critically examine assessment theories and practices, and if necessary adapt them to fit particular purposes, or develop new approaches that are consistent with our understanding of writing, reading, and teaching. Holistic scoring as traditionally defined came out of the measurement community and reinforces an approach to reading and writing that is acontextual and objective. Psychometric

theory, which is used to "validate" holistic scoring, assumes traits and abilities are normally distributed throughout the population, an assumption that is antithetical to what composition theory supports. These traits or abilities, according to traditional psychometrics are isolatable, quantifiable, and unchanging. Writing specialists, however, define writing as a contextual, communicative activity that is not transferable across time and place. Composition theory also assumes that writing "abilities" are influenced by instruction. These fundamental differences are significant and should not remain hidden or unarticulated but rather need to be addressed directly. By integrating experience and knowledge of composition, teaching, and psychometrics and confronting paradigmatic conflicts, Smith was able to create new approaches to assessment that honored composition scholarship and assessment demands.

Since the mid-1980s, there seems to be an accumulating body of composition research about placement (i.e., Borrowman , 1999; Decker, Cooper, & Harrington, 1993; Harrington, 1998; Haswell & Wyche-Smith, 1994; Huot, 1994; Lowe & Huot, 1997; Robertson, 1994; Royer & Gilles, 1998; Sommers, Black, Daiker, & Stygall, 1993). Unfortunately, the level of systematic and ongoing inquiry into these programs has been inconsistent, or at the very least inconsistently reported: Haswell and Wyche-Smith's (1994) work has developed into a comprehensive writing assessment system and a rich source of scholarship and ongoing research (e.g., Haswell, 1998, 2001; Haswell, Johnson-Shull, & Wyche-Smith, 1994; Haswell & McLeod, 1997). Yet other placement systems, such as the self-placement system used at Grand Valley State University (Royer & Gilles, 1998) or the small-group teaching model reported by Robertson (1994), provided very little rigorous research to support them and demonstrated lack of awareness of the complex assessment theories involved in designing and directing placement programs, but were legitimized through publication. Innovating and reconceptualizing placement can be important sources of knowledge, providing improved ways of meeting students' needs; however, without the appropriate inquiry, which demands an understanding of the complexity of the theories and assumptions informing writing and assessment practices, there is no way to justify revising or maintaining assessment procedures.

As a field, college composition has been quick to embrace new assessment practices—such as holistic scoring, portfolios, and directed self-placement—without sufficient understanding of the theories and assumptions that support them. When assessments are adopted and promoted without appropriate validation inquiry, we are not only jeopardizing our students' opportunities for learning and success—after all, writing assessments often function as institutional barriers—but we are ignoring a significant site of power and knowledge, undermining the legitimacy and professionalism of composition.

Notes

1. I have relied heavily on the published work of William L. Smith (1992, 1993) as well as numerous informal communications with him about his research at the University of Pittsburgh where he served as the Composition Program's director of testing for more than a decade. Bill not only responded to my never-ending questions, but he also read and commented on multiple drafts of this article.
2. I realize that many programs claim to use "holistic scoring," but my point is that often what is called holistic scoring is actually placement, as I explained earlier.
3. Understanding Smith's validation inquiry requires some sense of the University of Pittsburgh's Composition Program because this type of research is local and contextualized. Smith (1993) described Pitt's Composition Program as being based on four concepts:
   1. Writing is an effort to make meaning;
   2. Writing is closely related to reading;
   3. To make meaning, a writer must develop a sense of authority; and
   4. Students gradually come to a sense of authority.

   Consequently, in all of their courses, students respond to a sequence of assignments on a central topic (see Bartholomae, 1983; Coles, 1981; Bartholomae & Petrosky, 1987, for more detailed expositions of the basis for the program). It is important to note that composition courses were not considered "service" courses; consequently, students were not required to write research papers or papers in various modes (description, narration, etc).

   Because students have varied abilities along the four dimensions, the first-year composition program consisted of three courses, each addressing different writing problems and abilities. Course A was designed for students with serious problems with writing that indicate problems with reading and appropriating a text they have read. These students' essays lack development of ideas, lack coherence, are not well-organized, and do not address the issue. Commonly, these students inadequately summarize what they are asked to read or make general statements about the issue or topic, but they do not interrelate what they have read with their own ideas. These students also typically have patterns of surface level errors caused by their inability to proofread. Students who successfully complete this course take Course C.

   Course B is also designed for students who have significant writing problems such as coherence, organization, or development of ideas, but these problems are not related to their ability to read. Instead, they indicate a lack of a sense of text and a lack of authority. Surface error is common in their texts, typically caused by their lack of a sense of text. If asked whether they read their own texts as they read other ones, they will say they do not, and if pressed for reasons, they will say that their own reading does not merit such reading. Students pass from this class into Course C.

   Course C is designed for students who have the ability to read and make meaning but need more experience in developing their abilities, particularly

in dealing with problematic texts and in using writing as a means for working their way through complex problems. Some students are exempted from any composition course because the writing ability they demonstrate suggests that these courses would not be of significant value to them. (pp.144-145)

4. There are three published articles about this research (Smith, 1992, 1993; Smith et al., 1985). Much of the research went unpublished, although "Assessing the Reliability and Adequacy of Holistic Scoring" (Smith, 1993) reports in detail on several years worth of research focused on raters. In addition, Smith, often with graduate students, presented several conference papers about this placement testing research.
5. Part of this research was reported by Smith et al., 1985.

REFERENCES

Bartholomae, D. (1983). Writing assessments: Where writing begins. In P.L. Stock (Ed.), Forum: *Essays on theory and practice in the teaching of writing* (pp. 300-312). Upper Montclair, NJ: Boynton/Cook.

Bartholomae, D., & Petrosky, A. (1987). *Ways of reading: An anthology for writers*. New York: Bedford/St. Martin's Press.

Beck, A., Dautermann, J., Miller, C., Murray, K., & Powell, P. R. (1997). *The best of Miami University's portfolios 1997*. Miami, OH: Department of English, Miami University.

Borrowman, S. (1999). The trinity of portfolio placement: Validity, reliability, and curriculum reform. *WPA: Journal of Writing Program Administration, 1/2*, 7-27.

Cherry, R. D., & Meyer, P. R. (1993). Reliability issues in holistic assessment. In M.M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 109-141). Cresskill, NJ: Hampton Press.

Coles, W. E., Jr. (1981). *Composing II: Writing as a self-creating process*. Rochelle Park, NJ: Hayden Book.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer (Ed.), *Test validity* (pp. 3-17). Hillside: Erlbaum.

Decker, E., Cooper, G., & Harrington, S. (1993). Crossing institutional boundaries: Developing an entrance portfolio assessment to improve writing instruction. *Journal of Teaching Writing 12*, 83-104.

Harrington, S. (1998). New visions of authority in placement test rating. *WPA: Journal of Writing Program Administrator, 22*, 53-84.

Haswell, R. (1998). Multiple inquiries in the validation of writing tests. *Assessing Writing, 5*, 89-109.

Haswell, R. H. (Ed.). (2001). *Beyond outcomes: Assessment and instruction within a university writing program*. Westport, CT: Ablex.

Haswell, R., Johnson-Shull, L., & Wyche-Smith, S. (1994). Shooting Niagara: Making portfolio assessment serve instruction at a state university. *WPA: Journal of Writing Program Administration, 18*, 44-54.

Haswell, R., & McLeod, S. (1997). WAC assessment and internal audiences: A dialogue. In K. B. Yancey & B. Huot (Eds.), *Assessing writing across the curriculum: Diverse approaches and practices* (pp. 217-236). Greenwich, CT: Ablex.

Haswell, R., & Wyche-Smith, S. (1994). Adventuring into writing assessment. *College Composition and Communication, 45*, 220-236.

Huot, B. (1994). A survey of college and university writing placement practices. *WPA: Journal of Writing Program Administration, 17*, 49-65.

Huot, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication, 47*, 549-566.

Huot, B. (2002). *(Re)Articulating writing assessment for teaching and learning*. Logan: Utah State University Press.

Lowe, T. J., & Huot, B. (1997). Using KIRIS writing portfolios to place students in first-year composition at the University of Louisville. *Kentucky English Bulletin, 20*, 47-64.

Lutz, W. D. (1996). Legal issues in the practice and politics of assessment in writing. In E. M. White, W. D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practice* (pp. 33-44). New York: Modern Language Association.

Messick, S. (1989). Meaning and value in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*, 229-58.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5-12.

Moss, P. A. (1995). Themes and variations in validity theory. *Educational Measurement: Issues and Practices, 14*(2), 5-13.

Moss, P. A. (1998). Response: Testing the test of the test. *Assessing Writing, 5,* 111-122.

Murphy, S., Carlson, S., & Rooner, P. with the CCCC Committee on Assessment. (1993). *Report to the CCCC executive committee: Survey of postsecondary writing assessment practices*. Unpublished manuscript.

Pula, J. J., & Huot, B. A. (1993). A model of background influences on holistic raters. In M. M. Williamson & B. A. Huot (Eds.),*Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 237-265). Cresskill, NJ: Hampton Press.

Robertson, A. (1994). Teach not test: A look at a new writing placement procedure. *WPA: Journal of Writing Program Administration, 18*, 56-63.

Royer, D., & Gilles, R. (1998). Directed self-placement: An attitude of orientation. *College Composition and Communication, 50*, 54-70.

Shephard, L. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405-450.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5-8, 13, 24.

Smith, W. L. (1992). The importance of teacher knowledge in college composition placement testing. In R. J. Hayes (Ed.), *Reading empirical research studies: The rhetoric of research* (pp. 289-316) Norwood, NJ: Ablex.

Smith, W. L. (1993). Assessing the reliability and adequacy of placement using holistic scoring of essays as a college composition placement test. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 142-205). Cresskill, NJ: Hampton Press.

Smith, W. L. (1998). Introduction to Special Issue. *Assessing Writing, 5*, 3-6.

Smith, W. L., Hull, G. A., Land, Jr., R. E., Moore, M. T., Ball, C., Dunham, D. D., Hickey, L. S., & Ruzich, C. W. (1985). Some effects of varying the structures of a topic on college students' writing. *Written Communication 2*, 73-89.

Sommers, J., Black, L., Daiker, D. D., & Stygall, G. (1993). The challenges of rating portfolios: What WPAs can expect. *WPA: Journal of Writing Program Administrator, 17*, 7-29.

White, E. M. (1995). Apologia for the timed-essay. *College Composition and Communication, 46*, 30-45.

White, E. M. (1998). *Teaching and assessing writing* (2nd ed.). Portland, ME: Calendar Islands.

Williamson, M. M. (1993). An introduction to holistic scoring: The social, historical, and theoretical context for writing assessment. In M.M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 1-44). Cresskill, NJ: Hampton Press.

Wolcott, W. with Legg, S. M. (1998). *An overview of writing assessment: Theory, research, and practice*. Urbana, IL: National Council of Teachers of English.

Yancey, K. B. (1999). Looking back as we look forward: Historicizing writing assessment. *College Composition and Communication, 50*, 483-503.