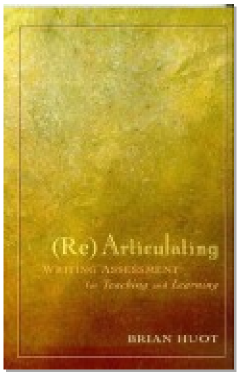
Rearticulating Writing Assessment for Teaching and Learning

Brian Huot



(Re) Articulating
Writing Assessment
for Teaching and Learning

BRIAN HUOT

Published by Utah State University Press

Huot, Brian.
Rearticulating Writing Assessment for Teaching and Learning.
Logan: Utah State University Press, 2002.
Project MUSE., https://muse.jhu.edu/.

4

TOWARD A NEW THEORY OF WRITING ASSESSMENT

Many writing teachers and scholars feel frustrated by, cut off from, or otherwise uninterested in the subject of writing assessment, especially assessment that takes place outside of the classroom for purposes of placement, exit, or program evaluation. This distrust and estrangement are understandable, given the highly technical aspects of much discourse about writing assessment. For the most part, writing assessment has been developed, constructed, and privatized by the measurement community as a technological apparatus whose inner workings are known only to those with specialized knowledge. Consequently, English professionals have been made to feel inadequate and naive by considerations of technical concepts like validity and reliability. At the same time, teachers have remained skeptical, and rightly so, of assessment practices that do not reflect the values important to an understanding of how people learn to read and write. It does not take a measurement specialist to realize that many writing assessment procedures have missed the mark in examining students' writing ability.

Many current debates about writing assessment issues (like whether or not to use standardized or local procedures for assessment, or whether or not we should abandon single-sample assessment in favor of portfolios) occur within a theoretical vacuum. Basically, we talk about and compare practices which have no articulated underlying theoretical foundation. Consequently, there are those scholars in composition who have questioned whether writing assessment is a theoretical enterprise at all. Anne Ruggles Gere (1980) suggested that writing assessment lacks a theoretical foundation; Faigley, Cherry, Jolliffe and Skinner (1985) elaborate this view by explaining that the pressing need to develop writing assessment procedures outstrips our ability to develop a theoretical basis for them.

My purpose in this chapter is to consider writing assessment from a theoretical perspective. By looking at the underlying principles which inform current practices, it is possible to consider how our beliefs and assumptions about teaching can and should inform the way we approach writing evaluation. I argue that, contrary to some scholarly opinion, writing assessment has always been a theory-driven practice. After tracing this theoretical thread to roots in classical test theory and its positivist assumptions, I illustrate how this theory has worked during the past couple of decades by examining current practices. These current practices and their underlying theoretical position are made all the more problematic if we consider the radical shift in testing theory that has been going on for the last two decades or so. This revolution in assessment theory has fostered performative approaches to assessment like the portfolio but, more importantly, it has actually redefined what it means for a test to be a valid measure of student ability. I express the need for the articulation of a new set of theoretical assumptions and practices for writing assessment. This theory will need to reconcile theoretical issues in measurement like validity and reliability with theoretical concerns in composition like rhetorical context and variable textual interpretations.

THEORETICAL FOUNDATION OF WRITING ASSESSMENT

Essentially, it is a mistake to assume that writing assessment has been developed outside the confines of a theoretical construct. While the field of composition often dates its birth in the 1960s, with the publication of *Research in Written Composition* or the convening of the Dartmouth Conference, work in writing assessment goes back several decades before that. Entrance examinations, implemented by Harvard and other universities before the turn of the twentieth century, were influenced during the 1920s by advances in educational testing brought on by the need to classify recruits for WWI, and were used to formalize writing assessment under the auspices of such testing institutions as the College Entrance Examination Board (CEEB). The development of writing assessment procedures, as we now know them, are the result of decades of research by the test development staff at CEEB and the Educational Testing Service (ETS). Building upon research started in the twenties (Hopkins 1921), the researchers at CEEB and ETS systematically established the procedures for writing assessment. (See Godshalk, Swineford and Coffman 1966 for a review of this literature.)

These efforts were undertaken under the auspices of classical test theory, which dictates that a measurement instrument has to be both valid and reliable. Classical test theory is based on a positivist philosophy which contends "that there exists a reality *out there*, driven by immutable natural laws" (Guba 1990, 19). Within the positivist foundation of classical test theory, it is possible to isolate a particular human ability, like writing, and measure it. Positivist reality assumes that student ability in writing, as in anything else, is a fixed, consistent, and acontextual human trait. Our ability to measure such a trait would need to recognize these consistencies and could be built upon psychometrics, a statistical apparatus devised for use in the social and hard sciences. Mathematics, as in physics, was conceived as the "language" of an empirical methodology that would assist in the discovery of fundamental laws governing human behavior. Guba (1990) labels this science "context free," because the laws revealed by this type of scientific method are held to be independent of the observer and the particular events in which they were discovered. Within such a paradigm, for example, the scores that students receive on a writing test like the National Assessment of Educational Progress (NAEP) are an accurate measure of the writing ability of the nation's students.[1] The results represent students' ability to write and can be compared from school to school and year to year, since psychometric methods ensure that their meaning exists outside of the context or time in which they were generated.

One of the reasons for writing scholars' belief that writing assessment is atheoretical stems from the fact that it was developed outside of the theoretical traditions that are normally considered part of composition. In addition to psychometricians at ETS, researchers trained to prepare secondary English teachers were responsible for much of the early work in writing assessment. Traditionally, an important aspect of the typical graduate program for English teachers is at least one or more courses in psychometric theory and practice. However, in the 1980s, as the study of writing became an interest for researchers trained in the humanities-based disciplines of rhetoric and composition (as opposed to the social sciences tradition of educational research), experimental and quantitative approaches to research became less important. Most researchers studying writing, without training in psychometric theory, were not aware of the theoretical origins of writing assessment. Most of them saw such concerns as the consistency (reliability) of assessment techniques simply as a matter of fairness (White 1993). Thus, issues that had originally been theoretical became pragmatic, and writing assessment became an apparently atheoretical endeavor.

The use of student writing to measure writing ability was unsupportable within classical test theory until the 1960s, because testing developers were unable until that time to devise methods for furnishing agreement among independent raters on the same paper. The theoretical foundation of writing assessment is apparent in our continuing emphasis on ensuring reliable methods for scoring student writing. Simply, interrater reliability has dominated writing assessment literature, a point I made in chapter two when referring to the overwhelming amount of research on reliability. As I noted, however, this trend has been changing a little during the past few years, as scholarship on writing assessment has begun to move beyond just establishing the procedures themselves. It is clear from even a cursory reflection on the history of direct writing assessment that not only were current methods for evaluation created within an established disciplinary framework, but that critical issues like reliability and validity existed and were defined within the context of classical test theory. The inability of scholars in composition to recognize the theoretical connections in writing assessment practices comes from the fact that it is a theory which has little familiarity or relevance for most people who teach and study the teaching of writing, especially at the college level. It is also possible to understand that our dissatisfaction with conventional means for assessing student writing (Broad 1994, 2000; Charney 1984; Faigley, Cherry, Jolliffe, and Skinner 1985; Gere 1980) has more to do with the theory that it informs it than with the practice of assessing student writing itself.

CURRENT PRACTICE IN WRITING ASSESSMENT

To be a viable option within classical test theory, writing assessment had to meet the same requirements expected of standardized tests. Conventional writing assessment's emphasis on uniformity and test-type conditions are a product of a testing theory that assumes that individual matters of context and rhetoric are factors to be overcome. From this perspective, a "true" measure of student ability can only be achieved through technical and statistical rigor. Most of the procedures and improvements in writing assessment have had as their goals either the reliability of the scoring or of the instrument itself. For example, writing assessment requires the development of writing prompts that are similar in difficulty and suitability for the testing population. Some early writing assessment programs produced great discrepancies in scores from one year to another because the writing tasks were of such variable difficulty (Hoetker 1982). Procedures for designing appropriate writing prompts often involve pilot testing and other measures (Ruth and Murphy 1988)[2] that ensure that students will perform fairly consistently on writing tasks used as part of the same or similar programs across different locations and times.

The bulk of writing assessment procedures are devoted to furnishing the raters with a means for agreement (Davis, Scriven, and Thomas 1987; Myers 1980; White 1994). Generally, raters are trained on a set of sample papers that are especially representative of particular scores on a scoring guideline or rubric. Once raters can agree consistently on scores for sample papers, they begin to score independently on “real, live papers.” Raters are periodically retrained or calibrated each day and throughout the scoring session(s) at appropriate intervals like after breaks for meals. These practices are consistent with a theory that assumes that teachers or other experts can identify good writing when they see it, and that in order for the assessment to be valid it must be consistent.

Within the positivist assumptions that construct and rely on the technology of testing, there is no need for different sets of procedures depending upon context, because writing ability is a fixed and isolated human trait, and this ability or quality can be determined though an analysis of various textual features. Depending on our purposes or resources, we can assess holistically for a general impression of quality, analytically for specific traits endemic to writing quality, or with a primary trait approach which treats rhetorical features of the writing assignment as the traits to be evaluated. Through the use of various scoring guidelines, we can decide what is of value within a student text and can base our judgments of a student’s writing upon differing approaches to that text. The assumption underlying these procedures is that writing quality exists within the text. See Figure 1 for a summary of the assumptions underlying traditional writing assessment procedures. While analytic (Freedman 1984) and primary trait (Veal and Hudson 1983) are usually considered a little better than holistic measures, holistic is cheapest and therefore considered the most popular (Veal and Hudson 1983).

Regardless of which form of writing assessment we choose to use, the emphasis is on the formal aspects of the procedures, the training of raters, the construction of scoring guidelines, the techniques necessary to guarantee interrater reliability. This emphasis is consonant with the importance of reliability in testing theory: “reliability is a necessary but not a sufficient condition for validity” (Cherry and Meyer 1993, 110). This importance for reliability has been adopted by college writing assessment

Figure 1

Traditional Writing Assessment
Procedures, Purposes and Assumptions

PROCEDURE	PURPOSE	ASSUMPTION
scoring guideline	recognize features of writing quality	writing quality can be defined and determined
rater training	foster agreement on independent rater scores	one set of features of student writing for which raters should agree
scores on papers	fix degree of writing quality for comparing writing ability and making decisions on that ability	student ability to write can be coded and communicated numerically
interrater reliability	calculate the degree of agreement between independent raters	consistency and standardization to be maintained across time and location
validity	determine that the assessment measures what it purports to measure	an assessment's value is limited to distinct goals and properties in the instrument itself

specialists and equated with fairness. Edward White provides a good summary of this position: "Reliability is a simple way of talking about fairness to test takers, and if we are not interested in fairness, we have no business giving tests or using test results" (1993, 93). Logically, then, the same procedures which ensure consistency should also provide fairness. However, this is not the case. First of all, we need to understand that reliability indicates only how consistent an assessment is. "Reliability refers to how consistently a test measures whatever it measures . . . a test can be reliable but not be valid" (Cherry and Meyer 1993, 110). For example, I could decide to measure student writing by counting the number of words in each essay (in fact a computer could count the words). This method could achieve perfect interrater reliability, since it is possible that two independent judges would count the same number of words for each paper. While reliable, we could hardly consider the method to be a fair evaluation of student writing. In order for an assessment instrument to be fair, we must know something about the nature of the judgment itself. Translating "reliability" into "fairness" is not only inaccurate, it is dangerous, because it equates the statistical consistency of the judgments being made with their value. While I applaud and agree with White's contention that writing assessment needs to be fair, and I agree that consistency is a component of fairness, there is nothing within current assessment procedures which addresses, let alone ensures, fairness.

Within the theory which currently drives writing assessment, the criteria for judging student writing are not an explicit part of assessment procedures. George Engelhard Jr., Belita Gordon and Stephen Gabrielson (1992) give us an example of a theoretically acceptable study of writing evaluation which contains some questionable criteria for assessing student writing. The study reports on the writing of 127,756 eighth-grade students and draws conclusions about the effects of discourse mode, experiential demand, and gender on writing quality. Three out of the five domains used for scoring all of this writing are "sentence formation," "usage" and "mechanics." The other two domains also emphasize the conventions of writing. "Content and organization" are relegated to one domain, with "clearly established controlling idea," "clearly discernible order of presentation" and "logical transitions and flow of ideas" as three of the six items in the domain. It is pretty easy to see how applicable these items are to the form of the standard five-paragraph essay. Domain number two, which is labeled "style," also focuses on the forms of writing. Although two of the items list "concrete images and descriptive language, [and] appropriate tone for topic, audience, and purpose," the other two are "easily readable [and] varied sentence patterns." While the study reports the results domain by domain, there is no attempt to differentiate the value of scores for content and organization over those for mechanics (1992, 320). What this research really reports is how the conventions and mechanics of student writing relate to the categories of analyses. This study might more easily and cheaply find out similar things about students by administering tests of grammar and mechanics with a question or two thrown in on thesis statements, topic sentences and transitions. However, the use of an essay test carries with it the weight or illusion of a higher degree of validity. Since the scoring of student writing follows recognizable procedures and produces acceptable levels of interrater reliability, there are no reasons under current traditional theories for assessment to question the study's results. Consequently, it was published in one of the profession's most prestigious journals.

My last example of current practices in writing assessment comes from a roundtable on reliability in writing assessment at a national convention during the 1990s. All of the presenters at the session were employed by testing companies. Two of the presenters (Joan Chikos Auchter 1993; Michael Bunch and Henry Scherich 1993) report using a set number of sample papers which had been given the same score by a large board of raters as their "true score or validity." Raters are trained to give the same score, and their suitability as raters depends upon their ability to match the score of the board. "The common characteristic of all of our readers is that they understand and accept the fact that they will score essays according to someone else's standards" (Bunch and Scherich 1993, 2). While such methods are effective in producing high interrater reliability, they are questionable even within psychometric theory. Validity is supposed to be separate from reliability, and here it is conflated for the set purpose of ensuring consistency in scoring. "True score" (for a good discussion relevant to writing assessment, see White 1994) is the score an examinee would get on a test if she could take it an infinite number of times; it would perfectly reflect her ability. The notion of true score used by these companies has nothing to do with a student's ability. Instead, the focus is directed to the scores she receives. True score becomes the number of scores the student gets on the same test, her ability forever fixed and accurate on one writing assignment because it is scored by many individuals. This last example of current practices in writing assessment is probably an extreme case, abusing the very theory that drives it. However, these practices are considered reputable and are used to make important educational decisions about students. One of these companies alone reports scoring three million student essays per year (Bunch and Scherich 1993, 3)—a number that has and will probably continue to increase, given the proliferation of state-mandated writing assessments in the 1990s and the impending federal assessments of public schools.

These practices and the theory that drive them are all the more lamentable when we consider that assessment theory has been undergoing a theoretical revolution during the last two decades, a revolution which has yet to filter down to the assessment of student writing. For example, in a report on the validity of the Vermont portfolio system, delivered at a national convention on measurement, the presenter elected to ignore more recent definitions of validity which also consider a test's influence on teaching and learning because "it would muddy the water" (Koretz 1993). Instead, he concentrated his remarks on the low interrater reliability coefficients and the consequently suspect validity of portfolio assessment in Vermont. What makes this type of scholarship in writing assessment even more frustrating is that portfolios are a form of performative assessment and are exactly the kind of practice that newer conceptions of validity are designed to support (Moss 1992). However, if we apply the more traditional, positivist notions of validity and reliability, we are judging a practice (portfolios, in this case) from outside the theoretical basis that informs it.

RECENT DEVELOPMENTS IN TESTING THEORY

It is necessary for those of us who teach writing and work in writing assessment to examine some of the radical shifts in testing theory which have been emerging, because these shifts have been influenced by the same philosophical and theoretical movements in the construction of knowledge that have influenced writing pedagogy. Some extreme positions call for the dismantling of validity itself, the cornerstone of classical test theory. For example, Guba and Lincoln (1989), in their book *Fourth Generation Evaluation*, posit a theory of evaluation based on the tenets of social construction in which validity is seen as just another social construct. Peter Johnston contends "that the term validity as it is used in psychometrics needs to be taken off life support" (1989, 510). Harold Berlak (1992) elaborates validity's insupportable position of privilege in testing.

> validity as a technical concept is superfluous. . . . I should point out that abandoning validity as a technical concept does not automatically mean abandoning all standardized and criterion referenced tests. It does mean, however, that [they] . . . may no longer be privileged as "scientific;" their usefulness and credibility are to be judged alongside any other form of assessment practice. (186)

These critiques of validity are critiques of a positivist notion of reality that assumes that human traits are distributed normally throughout the population and that these traits are distinct from the observer or tester and can thus be measured accurately across individual contexts. In fact, the power of psychometric procedures lies in their ability to render results that are accurate and generalizable to the population at large. These underlying positivist postures toward reality which inform traditional testing are partly responsible for the importance that objectivity and outside criteria for judging writing have in our thinking about the testing of writing. Judgments about student writing are often questioned as not being objective enough.

According to Johnston, the notion of objectivity in testing is linked to the positivist philosophy that has tightened the psychometric grip on educational testing.

> The search for objectivity in psychometrics has been a search for tools that will provide facts that are untouched by human minds. Classical measurement has enshrined objectivity in terms such as "objective" tests and "true score" (absolute reality) . . . The point is that no matter how we go about educational evaluation, it involves interpretation. Human symbol systems are involved, and thus there is no "objective" measurement. (1989, 510)

Johnston notes that even in the hard sciences the act of observation can alter what is being observed. For example, when light is used to view atomic particles, what we see is altered because of the effect of the light on what is observed. In his chapter on the history of writing assessment, Michael Williamson gives us another example which illustrates that reality is often an illusive quality even in science. He points out that an instrument like a telescope only gives the illusion of direct observation. "In fact, a telescope magnifies the light or radio waves reflected or emitted by cosmic bodies and does not result in direct observation at all" (Williamson 1993, 7).

In his discussion of objectivity, Johnston goes on to explain that in assessing students' abilities to read and write, interpretation plays an even larger role because communication depends on personal commitment, and texts cannot exist outside of the context and history in which they are produced. While those of us who teach writing have always known that we could only pretend to assess writing from an "objective" stance and therefore deferred to testing specialists for an objective view, Johnston contends that, "The search for objectivity may not simply be futile. I believe it to be destructive" (1989, 511). Drawing upon the work of Jerome Bruner, Johnston explains that if education is to create a change in individuals beyond the ability to regurgitate information, its focus cannot be "objective," because abilities like creativity, reflection, and critical thinking require a personal relationship with the subject. This negative influence of objectivity relates specifically to the assessment of writing, since good communication often requires the personal involvement of both writer and reader. The importance of reflection or point of view in writing is contradictory to an objective approach, because to assume a particular position is to be subjective (Johnston 1989, 511). New movements in testing theory which question the advisability of devising objective tests and maintaining equally objective evaluations of student performance have important implications for writing assessment, since those of us who teach and research literacy have always known that writing assessment could never be totally objective, and that writing which approached such objectivity would not be effective communication.

Although the diminishing need for objectivity will have an important effect on writing assessment, the biggest change will eventually be felt in developing notions of validity. For several decades, stretching back to the 1950s, validity has come to be defined as more than just whether or not a test measures what it purports to measure. Samuel Messick (1989a; 1989b) and Lee Cronbach (1988; 1989), two of the most prominent scholars of validity theory, revised their views throughout the 1980s. For Messick, validity is "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (1989b, 5). In this definition, there are two striking differences from traditional notions of validity. First of all, Messick includes multiple theoretical as well as empirical considerations. In other words, in writing assessment, the validity of a test must include a recognizable and supportable theoretical foundation as well as empirical data from students' work. Second, a test's validity also includes its use. Decisions based upon a test that, for example, is used for purposes outside a relevant theoretical foundation for the teaching of writing would have a low, unacceptable degree of validity. Cronbach's stance is similar. For Cronbach validity "must link concepts, evidence, social and personal consequences, and values" (1988, 4).

In both of these definitions of validity, we are asked to consider more than just empirical or technical aspects of the way we assess. In writing assessment, the technical aspects of creating rubrics, training raters, developing writing prompts, and the like have been the reasons why outside objective measures were superior to just having teachers read and make specific judgments about student writers. These new conceptions of validity question our preoccupation with the technical aspects of writing assessment procedures. In Cronbach's terms, we will need to link together these technical features with what we know about writing and the teaching of writing. In addition to establishing and expanding the theoretical and empirical foundation for assessing writing, both Messick and Cronbach's definitions require us to establish a theoretical foundation for the way we assess and to ensure that the evaluation of writing only be used for educational purposes which encourage the teaching and learning of writing.

Few important or long lasting changes can occur in the way we assess student writing outside of the classroom unless we attempt to change the theory which drives our practices and attitudes toward assessment. At present, assessment procedures that attempt to fix objectively a student's ability to write are based upon an outdated theory supported by an irrelevant epistemology. Emergent ideas about measurement define teaching, learning, and assessment in new ways, ways that are compatible with our own developing theories about literacy, though for the most part they have yet to filter down to the assessment of student writing. The result has been a stalemate for writing assessment. Although we were able to move from single-sample impromptu essays to portfolios in less than thirty years, we are still primarily concerned with constructing scoring guidelines and achieving high levels of interrater reliability.

EXAMINING AND UNDERSTANDING NEW PROCEDURES

This section explores our ability to construct a theory of writing assessment based upon our understandings about the nature of language, written communication, and its teaching. The bases for this theoretical exploration are current practices at universities who have been using assessment procedures unsupported by conventional writing assessment's reliance on the positivist, epistemological foundations of classical test theory. These new procedures recognize the importance of context, rhetoric, and other characteristics integral to a specific purpose and institution. The procedures are site-based, practical, and have been developed and controlled locally. They were created by faculty and administrators at individual institutions to solve specific assessment needs and to address particular problems. Individually, these procedures for assessing writing provide solutions for specific institutions. It is my hope to connect these procedures through their common sets of beliefs and assumptions to create the possibility of a theoretical umbrella. This theorizing can help other institutions create their own procedures that solve local assessment problems and recognize the importance of context, rhetoric, teaching, and learning. By themselves, each of these institutions has had to develop and create its own wheel; together, they can aid others in understanding the nature of their assessment needs and to provide solutions that "link together" the concerns of a variety of stakeholders.[3]

One of the most common forms of writing assessment employed by many institutions is the placement of students into various writing courses offered by a specific college or university. Traditionally, schools have used holistic scoring procedures to place students, adapting specific numerical scores, usually the combined or sum scores of two raters, to indicate placement for a particular class. Some of the earliest and most interesting procedures developed outside the traditional theoretical umbrella for writing assessment involve placement. Current traditional placement procedures require the additional steps necessary to code rater decisions numerically and to apply these numbers to specific courses. Research indicates that traditional procedures might be even more indirect, since talk-aloud protocols of raters using holistic methods for placement demonstrate that raters often first decide on student placement into a class and then locate the appropriate numerical score that reflects their decision (Huot 1993; Pula and Huot 1993). Newer placement programs end this indirection by having raters make placement decisions directly.

The first and most rigorously documented of the new placement programs was developed by William L. Smith (1993) at the University of Pittsburgh. His method involved using instructors to place students in specific classes based upon the writing ability necessary for success in the courses those instructors actually taught. This method of placing students proved to be more cost-efficient and effective than conventional scoring methods (Smith 1993). Such a placement program circumvents many of the problems found in current placement testing. Raters are hired in groups of two to represent each of the courses in which students can be placed. These pairs of raters are chosen because their most immediate and extensive teaching experience is in a specific course. A rater either decides that a student belongs in her class or passes the paper on to the rater for the class in which she thinks the student belongs. Using standard holistic scoring methods to verify this contextual placement scoring procedure, Smith found that students were placed into courses with greater teacher satisfaction and without the need for rubrics, training sessions, quantification, and interrater reliability.

While this method has been revised as the curriculum it supports is also revised (Harris 1996), these changes are in keeping with the local nature of this and other emergent writing assessment methods. Unlike traditional methods that centralize rating guidelines or other features of an assessment scheme, these site-based procedures can and should be constantly revised to meet the developing needs of an institution. For my purposes in this chapter, Smith's (1993) or other procedures that have been developed outside of a psychometric framework are less important for the utilization of the procedures themselves and more for their ability to define a set of principles capable of solving particular assessment problems, developed and revised according to local assessment needs.[4]

Another placement procedure, dubbed a two-tier process, has been developed at Washington State University, in which student essays are read by a single reader who makes one decision about whether or not students should enter the most heavily enrolled first-year composition course (Haswell and Wyche-Smith 1994). Students not so placed by the first-tier reader have their essays read in mutual consultation by a second tier of raters, experts in all courses in the curriculum. In this method, sixty percent of all students are placed into a course on the first reading. [5]

Pedagogically, these contextualized forms of placement assessment are sound because teachers make placement decisions based upon what they know about writing and the curriculum of the courses they teach. Placement of students in various levels of composition instruction is primarily a teaching decision. Smith (1993) analyzed the talk-aloud protocols of his raters and found that they made placement decisions upon whether or not they could "see" a particular student in their classrooms. Judith Pula and I (1993) report similar findings from interviewing raters reading placement essays in holistic scoring sessions. Raters reported making placement decisions not upon the established scoring guidelines on a numerical rubric but rather on the "teachability" of students. The context for reading student writing appears to guide raters regardless of rubrics or training found in many assessment practices (Huot 1993; Pula and Huot 1993).

While the first two procedures I've discussed have to do with placement, the others involve exit exams and program assessment. Michael Allen (1995) discusses his and his colleagues' experience with reading portfolios from various institutions. Allen found that readers who knew the context and institutional guidelines of the school at which the portfolios were written could achieve an acceptable rate of interrater reliability by just discussing the essays on-line over the internet, without any need for scoring guidelines or training sessions. Allen theorizes that readers are able "to put on the hat" of other institutions because they are experts in reading student writing and teaching student writers.

While Allen (1995) discusses the results and implications of reading program portfolios with a group of teachers across the country, Durst, Roemer, and Schultz (1994) write about using portfolios read by a team of teachers as an exit exam at the University of Cincinnati to determine whether or not students should move from one course to another. What makes their system different is that these "trios," as the three-teacher teams are called, not only read each others' portfolios but discuss that work to make "internal struggles [about value and judgment] outward and visible" (286). This system revolves around the notion that talk is integral to understanding the value of a given student portfolio. While White (1994) and Elbow and Belanoff (1986) have noted that bringing teachers together to talk about standards and values was one of the most important aspects of writing assessment, Durst, Roemer, and Schultz (1994) make the conversation between teachers the center of their portfolio exit scheme. They assert that their system for exit examination has benefits beyond the accurate assessment of student writing: "portfolio negotiations can serve as an important means of faculty development, can help ease anxieties about grading and passing judgment on students' work, and can provide a forum for teachers and administrators to rethink the goals of a freshman English program" (287). This public discussion of student work not only furnishes a workable method to determine the exit of particular students but also provides real benefits for the teachers and curriculum at a specific institution as newer conceptions of validity advocate (Cronbach 1988; 1989; Messick 1989a; 1989b; Moss 1992).

While all of the methods we have examined have distinctions predicated upon the context of their role(s) for a specific institution or purpose, they also share assumptions about the importance of situating assessment methods and rater judgment within a particular rhetorical, linguistic and pedagogical context. The focus of each of these programs is inward toward the needs of students, teachers and programs rather than outward toward standardized norms or generalizable criteria. In sharp contrast to the acontextual assumptions of traditional procedures (see figure one), these developing methods depend on specific assessment situations and contexts. Figure two summarizes the procedures and purposes of these emergent assessment methods.

IMPACT ON RELIABILITY

All of the procedures and the assumptions they hold either bypass or make moot the most important feature of current traditional writing assessment—the agreement of independent readers, or interrater reliability. Although Smith's (1993) procedures involve raters reading independently (without discussion or collaboration), rater agreement, by itself, is not crucial,

Figure 2

New, Emergent Writing Assessment Procedures, Purposes and Assumptions

Procedure	Purpose	Assumptions
raters from specific courses place students into their courses	writing placement	placement is a teaching decision based on specific curricular knowledge
one rater reads all essays and places 60% of all students; other 40% placed by expert team of consultants	writing placement	placement largely a screening process; teachers recognize students in primary course
rater groups discuss portfolios for exit or specific level of achievement	exit and program assessment	discussion and multiple interpretation necessary for high stakes decisions about students or programs
validity	determine accuracy assessment and impact of process on teaching and learning for a specific site and its mission and goals	value of an assessment can only be known and accountable to a specific context

because all raters are not equally good judges for all courses. Those decisions by the teachers of the course are privileged, since they are made by the experts for that course and that educational decision.

One of the possible reasons why we have historically needed methods to ensure rater agreement stems from the stripping away of context, common in conventional writing assessment procedures to obtain objective and consistent scores. This absence of context distorts the ability of individuals who rely on it to make meaning. For example, the most famous study involving the inability of raters to agree on scores for the same papers conducted by Paul Diederich, John French, and Sydell Carlton (1961) gave readers no sense of where the papers came from or the purpose of the reading. Given the total lack of context within which these papers were read, it is not surprising that they were scored without consistency. The absence of context in traditional writing assessment procedures could be responsible for the lack of agreement among raters that these procedures, ironically, are supposed to supply. The traditional response to raters' inability to agree has been to impose an artificial context, consisting of scoring guidelines and rater training in an attempt to "calibrate" human judges as one might adjust a mechanical tool, instrument or machine. White (1994) and other early advocates of holistic and other current traditional procedures for evaluating writing likened these scoring sessions to the creation of a discourse community of readers. However, Pula's and my (1993) study of the influence of teacher experience, training, and personal background on raters outlines the existence of two discourse communities in a holistic scoring session: one, the immediate group of raters, and the other, a community whose membership depends upon disciplinary, experiential and social ties. It seems practically and theoretically sound that we design schemes for assessment on the second discourse community instead of attempting to superimpose one just for assessment purposes.

Clearly, this inability of raters to agree in contextually stripped environments has fueled the overwhelming emphasis on reliability in writing assessment. Michael Williamson (1994) examines the connection between reliability and validity in writing assessment by looking at the ways more reliable measures like multiple-choice exams are actually less valid for evaluating student writing. Looking at validity and reliability historically, Williamson concludes that "the properties of a test which establish its reliability do not necessarily contribute to its validity" (1994, 162). Williamson goes on to challenge the traditional notion that reliability is a precondition for validity: "Thus, comparatively high reliability is neither a necessary nor a sufficient condition for establishing the validity of a measure" (1994, 162).

While Williamson contends that reliability should be just one aspect of judging the worthwhile nature of an assessment, Pamela Moss (1994) asks the question in her title, "Can There be Validity Without Reliability?" Moss asserts that reliability in the psychometric sense "requires a significant level of standardization [and that] this privileging of standardization is problematic" (1994, 6). Moss goes on to explore what assessment procedures look like within a hermeneutic framework. She uses the example of a faculty search in which members of a committee read an entire dossier of material from prospective candidates and make hiring decisions only after a full discussion with other members of the committee. In a later article, Moss (1996) explores the value of drawing on the work and procedures from interpretive research traditions to increase an understanding of the importance of context in assessment. Instead of interchangeable consistency within an interpretive tradition, reliability becomes a critical standard with which communities of knowledgeable stakeholders make important and valid decisions.

Interpretive research traditions like hermeneutics support the emerging procedures in writing evaluation because they "privilege interpretations from readers most knowledgeable about the context of assessment" (Moss 1994a, 9). An interpretive framework supports the linguistic context within which all writing assessment should take place, because it acknowledges the indeterminacy of meaning and the importance of individual and communal interpretations and values. Interpretive research traditions hold special significance for the assessment of student writing, since reading and writing are essentially interpretive acts. It is a truism in current ideas about literacy that context is a critical component in the ability of people to transact meaning with written language. In composition pedagogy, we have been concerned with creating meaningful contexts in which students write. A theory of assessment that recognizes the importance of context should also be concerned with creating assessment procedures that establish meaningful contexts within which teachers read and assess. Building a context in which writing can be drafted, read, and evaluated is a step toward the creation of assessment practices based on recognizable characteristics of language use. Assessment procedures that ignore or attempt to overcome context distort the communicative situation. Michael Halliday asserts that "Any account of language which fails to build in the situation as an essential ingredient is likely to be artificial and unrewarding" (1978, 29). Halliday's contention that "*All* language functions in contexts of situations and is relatable to those contexts" (1978, 32) is part of a consensus among scholars in sociolinguistics (Labov 1980), pragmatics (Levinson 1983), discourse analysis (Brown and Yule 1983), and text linguistics (de Beaugrande and Dressler 1981) about the preeminence of context in language use.

CREATING NEW ASSESSMENTS OF WRITING

Research on the nature of raters' decisions (Barritt, Stock and Clark 1986; Pula and Huot 1993) indicate the powerful tension teachers feel between their roles as readers and raters in an assessment environment. An appropriate way to harness this tension is to base assessment practices within specific contexts, so that raters are forced to make practical, pedagogical, programmatic, and interpretive judgments without having to define writing quality or other abstract values which end up tapping influences beyond the raters or test administrators' control. As Smith (1993) and Haswell and Wyche-Smith (1994) have illustrated with placement readers, Durst, Roemer, and Schultz (1994) with exit raters, and Allen (1995) with program assessment, we can harness the expertise and ability of raters within the place they know, live, work and read. Assessment practices need to be based upon the notion that we are attempting to assess a writer's ability to communicate within a particular context and to a specific audience who needs to read this writing as part of a clearly defined communicative event.

It follows logically and theoretically that rather than base assessment decisions on the abstract and inaccurate notion of writing quality as a fixed entity[6]—a notion which is driven by a positivist view of reality—we should define each evaluative situation and judge students upon their ability to accomplish a specific communicative task, much like the basic tenets of primary trait scoring. However, instead of just basing the scores upon rhetorical principles, I propose that we design the complete assessment procedure upon the purpose and context of the specific writing ability to be described and evaluated. The three major means for assessing writing, holistic, analytic and primary trait, are largely text-based procedures which merely manipulate the numerically-based scoring guidelines. These procedures would be replaced by contextually and rhetorically defined testing environments. The type of scoring would be identified by the genre of the text to be written, the discipline within which it was produced and the type of decisions the raters are attempting to make.

In business writing, for example, students might be required to condense extensive documents into a few paragraphs for an executive summary. Students in the natural or physical sciences might be given the data obtained through research procedures and be required to present such information in a recognizable format, complete with applications. In environmental writing, where speed and the ability to synthesize technical information for a lay audience is crucial, students might be given a prompt they have never seen and be asked to produce text in a relatively short period of time. Instead of current methods, we would have placement testing for first-year composition or business competency writing or high school exit writing in which the purpose, context and criteria would be linked together to create procedures built upon the rhetorical, linguistic, practical and pedagogical demands of reading and writing in a specific context. Debates, for example, about the use of single-samples or portfolios (Purves 1995; White 1995a; 1995b), would be moot, since the number and type of writing samples and the method for producing the texts would depend upon the specific assessment context. The criteria for judgment would be built into a method and purpose for assessment and would be available, along with successful examples of such writing to the student writers. Not only do these proposed methods for assessing writing reject scoring guidelines, rater training for agreement, calculations of interrater reliability, and the other technologies of testing, but they also connect the context, genre, and discipline of the writing with those making evaluative decisions and the criteria they use to judge this writing. When we begin to base writing evaluation on the context of a specific rhetorical situation adjudged by experts from within a particular area, we can eliminate the guessing students now go through in preparing for such examinations, as well as the abstract debates and considerations about the best procedures for a wide variety of assessment purposes.

TOWARD A NEW THEORY OF WRITING ASSESSMENT

The proposed writing assessments we have discussed and other procedures like them exist outside the "old" theoretical tenets of classical test theory.[7] Instead of generalizability, technical rigor and large scale measures that minimize context and aim for a standardization of writing quality, these new procedures emphasize the context of the texts being read, the position of the readers and the local, practical standards that teachers and other stakeholders hold for written communication. There is a clear link between the judgments being made and the outcome of these judgments that is neither hidden nor shaded by reference to numerical scores, guidelines or statistical calculations of validity or reliability. These site-based, locally-driven procedures for evaluating student writing have their roots in the methods and beliefs held by the teachers who teach the courses that students are entering or exiting, or in the program under review. In this light, there is a much clearer connection between the way writing is taught and the way it is evaluated. For the last two or three decades, writing pedagogy has moved toward process-oriented and context-specific approaches that focus on students' individual cognitive energies and their socially positioned identities as members of culturally bound groups. In contrast, writing assessment has remained a contextless activity emphasizing standardization and an ideal version of writing quality.

These emergent methods can be viewed under a new theoretical umbrella, one supported by evolving conceptions of validity that include the consequences of the tests and a linking of instruction and practical purposes with the concept of measuring students' ability to engage in a specific literacy event or events. These procedures also have their bases in theories of language and literacy that recognize the importance of context and the individual in constructing acceptable written communication. These methods are sensitive to the importance of interpretation inherent in transactional and psycholinguistic theories of reading. Although it is premature to attempt any overall or complete discussion of the criteria for newer conceptions of writing assessment, figure three provides a set of preliminary principles extrapolated from our consideration and discussion of these new assessment procedures and their connection to current theories of measurement, language, and composition pedagogy. Like the assessment practices themselves, any writing assessment theory will need to be considered a work in progress as new procedures and the theories that inform them continue to advance our theoretical and practical understanding of writing assessment.

Figure 3

Principles For a New Theory
And Practice of Writing Assessment

Site-Based

An assessment for writing is developed in response to a specific need that occurs at a specific site. Procedures are based upon the resources and concerns of an institution, department, program or agency and its administrators, faculty, students or other constituents.

Locally-Controlled

The individual institution or agency is responsible for managing, revising, updating and validating the assessment procedures, which should in turn be carefully reviewed according to clearly outlined goals and guidelines on a regular basis to safeguard the concerns of all those affected by the assessment process.

Context-Sensitive

The procedures should honor the instructional goals and objectives as well as the cultural and social environment of the institution or agency and its students, teachers and other stakeholders. It is important to establish and maintain the contextual integrity necessary for the authentic reading and writing of textual communication.

Rhetorically-Based

All writing assignments, scoring criteria, writing environments and reading procedures should adhere to recognizable and supportable rhetorical principles integral to the thoughtful expression and reflective interpretation of texts.

Accessible

All procedures and rationales for the creation of writing assignments, scoring criteria and reading procedures, as well as samples of student work and rater judgment, should be available to those whose work is being evaluated.

Developing writing assessment procedures upon an epistemological basis that honors local standards, includes a specific context for both the composing and reading of student writing and allows the communal interpretation of written communication is an important first step in furnishing a new theoretical umbrella for assessing student writing. However, it is only a first step. We must also develop procedures with which to document and validate their use. These validation procedures must be sensitive to the local and contextual nature of the procedures themselves. While traditional writing assessment methods rely on statistical validation and standardization that are important to the beliefs and assumptions that fuel them, developing procedures will need to employ more qualitative and ethnographic validation procedures like interviews, observations and thick descriptions to understand the role an assessment plays within a specific program or institution. We can also study course outcomes to examine specific assessments based upon specific curricula. William L. Smith's (1993) validation procedures at the University of Pittsburgh and Richard Haswell's (2001) at Washington State can probably serve as models for documenting emerging procedures.

These local procedures can be connected beyond a specific context by public displays of student work and locally developed standards. Harold Berlak (1992) proposes that the use of samples from several locations be submitted to a larger board of reviewers who represent individual localities and that this larger board conduct regular reviews of student work and individual assessment programs. Pamela Moss (1994a) outlines a model in which representative samples of student work and localized assessment procedures work can be reviewed by outside agencies. Allen's (1995) study furnishes a model for a "board" of expert readers from across the country to examine specific assessment programs, including samples of student work and the local judgments given that work.[8] His use of electronic communication points out the vast potential the Internet and the Web have in providing the linkage and access necessary to connect site-based, locally controlled assessment programs from various locations. As Moss

(1994a) cautions, we have only begun to revise a very established measurement mechanism, and there is much we still need to learn about how to set up, validate and connect local assessment procedures.

It is important to note that all of the procedures I have highlighted as depending upon an emergent theory of assessment that recognize context and local control were developed at the college level. Even state-mandated portfolio systems like those in Kentucky and Vermont continue to be standardized in order to provide for acceptable rates of interrater reliability. It is imperative that we at the college level continue our experimentation and expand our theorizing to create a strong platform for new writing assessment theory and practice. Connecting those who work in college writing assessment with those engaged in writing assessment from the educational measurement community, as I advocate in chapter two, can not only foster a more unified field but can also provide the possibility of rhetorical and contextual writing assessment for all students. We need to begin thinking of writing evaluation not so much as the ability to judge accurately a piece of writing or a particular writer, but as the ability to describe the promise and limitations of a writer working within a particular rhetorical and linguistic context.

As much as these new procedures for writing assessment might make practical and theoretical sense to those of us who teach and research written communication, they will not be widely developed or implemented without much work and struggle, without an increased emphasis on writing assessment within the teaching of writing at all levels. English teachers' justifiable distrust of writing assessment has given those without knowledge and appreciation of literacy and its teaching the power to assess our students. The ability to assess is the ability to determine and control what is valuable. Standardized forms of assessment locate the power for making decisions about students with a central authority. Harold Berlak (1992) labels the educational policies of the Reagan-Bush era "incoherent," because while policy makers called for increased local control of schools, they also instituted massive standardized testing, rendering any kind of local decision-making superfluous. Changing the foundation that directs the way student writing is assessed involves altering the power relations between students and teachers, and teachers and administrators. It can also change what we will come to value as literacy in and outside of school. At this point, the door is open for real and lasting changes in writing assessment procedures. We who teach and research written communication need to become active in assessment issues and active developers of these new, emergent practices. In the past, current writing assessment procedures were largely developed by ETS and other testing companies outside of a community of English or composition teachers and were based upon a set of assumptions and beliefs irrelevant to written communication. Unlike the past, it is time for us to go through the door and take charge of how our students are to be evaluated. It is time to build and maintain writing assessment theories and practices which are consonant with our teaching and research.