Brian Huot

Toward a New Theory of Writing Assessment

Many composition 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.

At the core of this inability to communicate are basic theoretical differences between the measurement and composition communities (White, “Language”). Writing assessment procedures, as they have been traditionally constructed, are designed to produce reliable (that is, consistent) numerical scores of individual student papers from independent judges. Traditional writing assessment practices are based upon classical test theory, with roots in a positivist epistemology that assumes “that there exists a reality out there, driven by immutable natural laws” (Guba 19). The assumption is that student ability in writing, as in anything else, is a fixed,

Brian Huot teaches graduate and undergraduate courses on writing at the University of Louisville, where he directs the composition program. He is coeditor of Assessing Writing, the only journal devoted to writing assessment. This essay culminates several years of thinking, talking, and writing about the need to articulate the theories behind writing assessment practices.
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 sciences.

Within such a paradigm, for example, the scores students receive on a writing test like the National Assessment of Educational Progress 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. In such large scale assessment, individual matters of context and rhetoric are to be overcome in favor of producing a "true" measure of student ability whose validity can only be established through technical and statistical rigor. These beliefs and assumptions put enormous faith in the technology of testing, things like the development of scoring guidelines or rubrics, the training of raters, the scores papers receive, and the statistical calculation of interrater reliability. It is through this psychometric technology that we can isolate the "reality" —in this case student writing ability—which a positivist epistemology assumes is "out there." Table 1 provides a summary of the procedures used in traditional writing assessment, the purpose of these procedures and the assumptions upon which they are based.

Currently, new ideas in measurement theory are being supported by the same theoretical movements in the social construction of knowledge that are used to explore and explain written communication in a postmodern age. At the heart of the movement in measurement theory towards considerations of context and individuality is the evolution of theories of validity (see Moss, "Shifting," for a good overview). Since the 1980s, validity has come to be defined as more than the traditional notion that a test measures what it purports to measure. Several validity scholars, prominent among them Samuel Messick and Lee Cronbach, have been revising their views for nearly 20 years.

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" (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. To be valid, writing assessment would need input from the scholarly literature about the teaching and learning of writing. Second, a test's validity also includes its use. A test, for example, which is used for purposes outside a relevant theoretical foundation for the teaching of writing would be an invalid measure of writing ability. Cronbach's stance
is similar. For Cronbach validity “must link concepts, evidence, social and personal consequences and values” (4). Under these terms, many current procedures now in use could be seriously challenged as invalid. Both of these newer definitions of validity question current-traditional preoccupations with the technical aspects of writing assessment procedures. Cronbach contends that we will need to link together assessment procedures with what we know about writing pedagogy and the impact of assessment procedures on teaching and learning. Both these new notions of validity 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.

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 which 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 which 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 have been able to move from single-sample impromptu essays to portfolios in less than 20 years, we are still primarily concerned with constructing scoring guidelines and achieving high rates of interrater reliability.

This essay 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 that have been using assessment procedures unsupported by conventional writing assessment's reliance on the positivist, epistemological foundations of classical test theory. Instead, 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 in order 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 to understand the nature of their assessment needs and to provide solutions that link together the concerns of a variety of "stakeholders."²

Examining and Understanding New Procedures

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. This makes sense because 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 often raters first decide on student placement into a class and then locate the appropriate numerical score that reflects their decision (Huot; Pula and Huot). Newer placement programs end this indirectness by having raters make placement decisions directly.

One of the first and most rigorously documented of the new placement programs was developed by William L. Smith 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). 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.

Recently, this method has been under revision as the curriculum it supports is also revised (Harris). This revision is 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. It might be best to think of the Smith method for placement as a prototype, since it requires a very stable pool of raters that teach specific courses on a consistent basis. In keeping with the purpose of this essay, Smith's 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.

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). 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, 60% of all students are placed into a course on the first reading.

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 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 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; Pula and Huot).

While the first two procedures I’ve discussed have to do with placement, the others involve exit exams and program assessment. Michael Allen 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. Borrowing a term Bill Condon used in his keynote at the Scottsdale Conference on “Portfolios, WAC and Program Assessment,” Allen offers the notion of “shared evaluation” to describe the experience of reading portfolios together. Such an evaluation would include:

(1) A tentative score based on an evaluator’s reading of a student text
(2) An openness of evaluators to other evaluators’ perspectives
(3) A rapid exchange of discursive analysis of the student text
(4) And examination of assessment issues that may arise from the exchange of evaluators’ perspectives (84)

While Allen discusses the results and implications of reading program portfolios with a group of teachers across the country, Durst, Roemer, and Schultz 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 differ-
ent 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 (Teaching) and Elbow and Belanoff 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 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; Messick; Moss, “Shifting”).

While each 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, these developing methods depend on specific assessment situations and contexts. Table 2 on page 556 summarizes the procedures and purposes of these emergent assessment methods.

Impact on Reliability

All of these procedures 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 procedures involve raters reading independently (without discussion or collaboration), rater agreement, by itself, is not crucial 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. In current traditional writing assessment agreement is “a necessary but not a sufficient condition for validity” (Cherry and Meyer 110). In other words, without a sufficient level of agreement between rat-
Table 2  
**New, Emergent Writing Assessment**  
Procedures, Purposes, and Assumptions

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Purpose</th>
<th>Assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raters from specific courses place students into their courses</td>
<td>Writing placement</td>
<td>Placement is a teaching decision based on specific curricular knowledge</td>
</tr>
<tr>
<td>One rater reads all essays and places 60% of all students; other 40% placed by expert team of consultants</td>
<td>Writing placement</td>
<td>Placement largely a screening process; teachers recognize students in primary course</td>
</tr>
<tr>
<td>Rater groups discuss portfolios for exit or specific level of achievement</td>
<td>Exit and program assessment</td>
<td>Discussion and multiple interpretation necessary for high stakes decisions about students or programs</td>
</tr>
<tr>
<td>Validity</td>
<td>Determine accuracy of assessment and impact of process on teaching and learning for a specific site and its mission and goals</td>
<td>Value of an assessment can only be known and accountable in a specific context</td>
</tr>
</tbody>
</table>

A writing assessment procedure can not be valid. For this reason, reliability has historically dominated the literature on writing assessment, and it was only after procedures such as scoring guidelines and rating training could guarantee agreement between raters that the reading and scoring of student writing (direct writing assessment) became psychometrically viable. Before that, only indirect measures that consist of multiple choice tests of usage and mechanics could be supported by classical test theory.

While the need for reliability is theoretically mandated within classical test theory, it has been supported on practical and ethical grounds in composition by those who see it as a means of ensuring fairness to students.
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” (“Holistic” 93). Logically, then, the same procedures which ensure consistency should also provide fairness. However, this is not the case. We should understand that in writing assessment interrater reliability means consistency among raters and nothing else. “Reliability refers to how consistently a test measures whatever it measures...a test can be reliable but not be valid” (Cherry and Meyer 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 my method to be a fair evaluation of student writing. Consistency is only one aspect of fairness. In order for an assessment instrument to be fair, we must know something about the nature of the judgment.3 Procedures that involve teachers in development and discussion and reflect clearly defined and negotiated local standards should provide for fair and responsible judgments of student writing. Translating reliability into fairness is not only inaccurate, it is dangerous because it equates statistical consistency with value about the nature of the judgments being made.

One of the possible reasons 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 was conducted by Paul Diederich, John French, and Sydell Carlton. Three hundred papers were distributed to 53 readers representing six different fields. Readers were given no sense of where the papers came from or the purpose of the reading. Given this lack of contextual cues, it is not surprising that 90% of the papers got at least seven different scores on a nine-point scale. The absence of context in traditional writing assessment procedures could be responsible for the lack of agreement among raters which these procedures are, ironically, 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 (Teaching) 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 and
Huot's 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.

This inability of raters to agree in contextually stripped environments has fueled the overwhelming emphasis on reliability in writing assessment. Michael Williamson 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 ways to evaluate 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" (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" (162).

While Williamson contends that reliability should be just one aspect of judging the worthwhile nature of an assessment, Pamela Moss 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" (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 more recent article, Moss ("Enlarging") 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. Within an interpretive tradition, reliability becomes not interchangeable consistency but rather 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" ("Can" 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 interpre-
tive 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 for students to write in. 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 procedures based on recognizable characteristics of language use. Assessment procedures which 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" (29). Halliday’s contention that "language functions in contexts of situations and is relatable to those contexts" (32) is part of a consensus among scholars in sociolinguistics (Labov), pragmatics (Levinson), discourse analysis (Brown and Yule), and text linguistics (de Beaugrande and Dressler) about the preeminence of context in language use.

Creating New Assessments of Writing

Research on the nature of raters’ decisions (Barrit, Stock, and Clark; Pula and Huot) indicate the powerful tension teachers feel between their roles as reader and rater 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 along with Haswell and Wyche-Smith have illustrated with placement readers, Durst, Roemer, and Schultz with exit raters, and Allen 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, 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 in which varying purposes, contexts, 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. Current debates, for example, about the use of single-samples or portfolios (Purves, "Apologia;" White, "Apologia" and "Response") 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 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. 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 teachers and other stakeholders establish 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 students are entering or exiting or 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. On the contrary, 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 reader response and psycholinguistic theories of reading. Although it is premature to attempt any full-blown discussion of the criteria for newer conceptions of writing assessment, figure 1 on page 562 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.

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 for 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 such assessment. These validation procedures must be sensitive to the local and contextual nature of the assessment being done. While traditional writing assessment methods rely on statistical validation and standardization that are important to the beliefs and assumptions that fuel them, emergent procedures will need to employ more qualitative and ethnographic validation procedures—like interviews, observations, and
SITE-BASED
An assessment for writing is developed in response to a 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 that should 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.

ACCESSIBILITY
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.

Figure 1: Principles For a New Theory And Practice of Writing Assessment

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. Smith's validation procedures at the University of Pittsburgh and Haswell's 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 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 ("Validity in High Stakes") out-
lines a model in which representative samples of student work and localized assessment procedures work can be reviewed by outside agencies. Allen’s 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. His use of electronic communication points out the vast potential the Internet and World Wide Web have in providing the linkage and access necessary to connect site-based, locally-controlled assessment programs from various locations. As Moss 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.

Inherent in a new conception of writing assessment is the strict limitation of what it can and cannot do. For example, large scale testing, which strives to define writing quality for a huge population of students, will have far less value than it now holds because context and population size will be part of a theoretical constraint on evaluation procedures, much in the same way most composition courses now impose some sort of enrollment cap. Instead, local results can be sampled and combined to replace current procedures that presently attempt to assess thousands of students across rhetorical and contextual boundaries. While Moss (“Enlarging”) and others offer the interpretive tradition as one possible way of supporting assessment procedures, it is important to remember that various epistemological stances are not neutral or innocent. For writing assessment, it is difficult to see the value in maintaining procedures based on a positivist epistemology. These new assessment schemes are context-rich and rely upon raters knowing as much as possible about the papers, the students, the purpose of the evaluation, the consequences of their decisions and the decisions of fellow raters. Within these procedures, the sacred cow of writing assessment, interrater reliability, becomes irrelevant because agreement on blind readings is no longer a crucial element for an accurate or valid evaluative decision. Instead, our attention is directed toward creating an assessment environment for reading and writing that is sensitive to the purpose and criteria for successful communication in which student ability in writing becomes part of a community’s search for value and meaning within a shared context.

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, so that we can see the emergence of rhetorical and contextual writing assessment for all students. We need to begin thinking of writing evaluation in new ways, not so much as the ability to judge accurately a piece of writing or a particular writer but to be able 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 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. Composition's justifiable distrust of writing assessment has given those outside of the discipline 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 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 which 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.

Notes

1. It is important to note that only a certain, "trend" sample of NAEP'S writing assessment claims to measure writing ability from year to year.
2. An assessment term that refers to all those who are affected by the measurement process, typically students, teachers administrators and parents. See Guba and Lincoln for a full treatment of the term and concept.
3. Current traditional procedures do not make the content of the scoring guideline an explicit part of the validation process. The
unfortunate result is that some "acceptable" assessments of student writing have problematic scoring criteria. For example, George Englehard Jr., Belita Gordon and Stephen Gabrielson report on having assessed the writing of over 125,000 students in which rater judgment was largely based on language conventions. While such assessment practices furnish the illusion of having evaluated student writing, in reality we have learned little more that what a multiple choice test might provide.

4. While I arrive at this idea theoretically, Alan Purves, in "Reflections on Research and Assessment in Written Composition," details the breakdown of writing quality as a concept in a study undertaken by the International Association of Educational Achievement on student writing in fourteen countries.

5. This movement away from psychometric procedures has been underway for some time (Barrit, Stock, and Clark; Carini; Faigley, Cherry, Jolliffe, and Skinner). There are many institutions employing similar, locally-developed procedures. SUNY Stony Brook, for example, has students write placement essays as part of a two-hour class on writing. The essay is read and judged by two teachers, one of which taught that group of students (Robertson). At the University of Louisville, teachers have met in groups to discuss and evaluate student portfolios as part of an evaluation of general education. We have adapted Smith's scheme to read high school portfolios for placement, and the English Department piloted a program last year in which teachers' portfolios were read collaboratively as part of an institutional evaluation of individual departments.

Works Cited


Harris, Joseph. Personal Correspondence, June, 1996.


