INTRODUCTION

This essay looks at the relationship between theory and practice in writing assessment. While this volume recognizes the many connections between theory, practice and research in composition, scholars in composition who work on writing assessment have not always shared an awareness of its theoretical assumptions and beliefs. Moreover, such recognition or awareness is crucial, we believe, if writing assessment is ever going to be an important component of composition as a discipline and in the practices of composition teachers and writing program administrators. The topic of theory and practice in writing assessment is especially important to us because a 1986 conversation about “holistic scoring being a practice without a theory” spawned a series of conversations, which in turn spawned three dissertations (one written by Brian and all three directed by Mike), several conference sessions, an edited collection, articles and book chapters we’ve written individually and together, and this essay. It’s interesting, then, for us to go back to the site of our initial interest and curiosity about writing assessment. It is a site that we have revisited in our work about writing assessment in a general attempt to excavate the theories, epistemologies, beliefs and assumptions that support much practice in writing assessment. Of course, the intersection of theory and practice is important beyond its personal significance, since we believe that to change the way writing assessment is designed and implemented, we must alter the way it is theorized and constructed as an intellectual space.

It’s worth noting that writing assessment procedures were not always recognized as even having an intellectual space. Nearly two decades ago, Anne Gere (1980) lamented the lack of an articulated set of principles for writing assessment. Five years later, Faigley, Cherry, Jolliffe and Skinner (1985) made a similar
statement based upon the same rationale. According to this line of reasoning, we had simply been too busy trying to assess writing to spend time developing a theory upon which our emergent practices might be based. We should note that in the early 1980s, holistic scoring and other direct forms of writing assessment (assessment that included the evaluation of student writing) were just emerging. While writing had been assessed regularly since the 1800s, the technology of direct writing assessment was formally developed in the 1960s after decades of work by the test developers at the College Entrance Examination Board (CEEB) and later by the Educational Testing Service (ETS).

As most of us interested in writing assessment are now aware, procedures developed by measurement professionals at ETS and CEEB outside the field of English Studies and the field of rhetoric and composition were supported by beliefs and assumptions from classical test theory within a positivist epistemology. Within such a belief system, there exist discernible truths which we can, if rigorous and objective enough in our methods, discover. The basis, then, of our first conversations about writing assessment proved groundless because holistic scoring has always been a theory-driven practice (Huot 1996; Williamson 1993). Such a discovery, however, merely took our conversations about theory and practice in new directions and provided new insights. For example, it became clear that the major so-called debate over the efficacy of indirect writing assessment (multiple choice tests) versus direct writing assessment was not really a debate between measurement specialists and teachers as it had been constructed (White 1993) because both sets of procedures aim to create the reliable measurement of student writing, and the achievement of that reliability was the central issue not only in this particular debate but within the central tenets of psychometric theory and the procedures this theory spawned. The debate over multiple choice tests versus essays was conducted by the measurement community over principles defined and articulated according to the theoretical basis of their intellectual community. Writing teachers were certainly interested in the outcome, but we were not active participants who advanced ideas and ultimately developed the procedures. This is an important point because some scholars in writing assessment (Scharton 1996; White 1994) advocate that composition teachers become more familiar with psychometric procedures, so that they can be taken more seriously in conversations and debates with the measurement community. While we will take up this particular point about the importance of scholars in rhetoric and composition learning statistical operations associated with a psychometric approach to writing assessment, we do believe that work in writing assessment needs to be interdisciplinary and to draw upon theories of teaching, literacy and validity in educational measurement. In this chapter we outline some basic tenets in validity theory because, as Pamela Moss points out, “the field of college writing assessment appears seriously isolated from the larger educational measurement community” (1998, 113). Our outline of validity and its implications for writing assessment permits us to look at competing models for writing assessment.
theoretically as we conclude the chapter with a comparison of the different the-
oretical and epistemological positions supporting such practices and an exami-
nation of the significance of these differences.

IMPLICATIONS OF VALIDITY THEORY FOR WRITING ASSESSMENT

Any interdisciplinary consideration of writing assessment needs to consider the
concept of validity in educational measurement. While the implications validity can have for studying writing assessment are extensive, validity is not a subject most of us in Composition are familiar with. We hope to provide a concise discussion on the nature of validity itself and on its evolution over the last several decades as a way not only of showcasing the interdisciplinary nature of writing assessment but of providing a useful set of ideas for understanding the way writing assessment can best work within the context of Composition. Like most theoretical constructs, validity continues to be redefined, debated and implemented within a changing world. This evolution is especially important for writing assessment because recent changes in validity mirror the many changes in the social construction of knowledge that have affected the way language and literacy are studied, theorized and taught in the late twentieth century. The literature on validity is extensive, dating back several decades, though its current theoretical orientation began in the early 1950s (Shephard 1993).

In its simplest terms, validity is concerned with how accurately an assessment hits a target. Reliability is concerned with the ability of an assessment to continue to hit the target. To be accurate over time, an assessment has to be reliable. That reliability is a necessary but insufficient condition for validity is an accepted maxim in traditional notions of writing and educational assessment. Thus, contemporary statistical procedures that examine the accuracy of an assessment consider its precision or reliability. Any deviations in reliability were considered to be sources of error, detracting from the overall validity of a measurement. These sources of errors were aggregated to produce mathematical calculations of validity. Initially, the literature in direct writing assessment was dominated by the issue of reliability although in the last few years this has begun to change as work in writing assessment has begun to move beyond the establishment of the procedures themselves. Traditional procedures for evaluating the validity of an assessment procedure are built upon an axiom that all data are distributed normally, that is, that they form the shape of a bell curve when plotted in geometric space. When data do not distribute normally, that is, when the range of the data is restricted in some way, the variance, a measure of how spread out the data are, is constrained. The more constrained the data, the lower the reliability. This phenomenon emerges from the mathematical fact that greater variance tends to produce greater reliability, all other things being equal. Thus, writing assessment procedures that create normal distributions of students are more reliable than those that create more narrow distributions. For example, a multiple choice test with a large number of items is going to
have a potentially larger range of scores than a holistically scored essay that only has four score points. Validation processes are designed to help tests create a normally distributed set of scores. Items are selected for a test precisely in order to insure that students are spread out over a range that is close to a bell curve. Thus, the mathematical axiom becomes a self fulfilling prophecy. Because a normal distribution is assumed a natural phenomenon in mathematical theory, tests are constructed to uncover that characteristic. Consequently, because tests are believed to be more valid when they are more reliable, they are constructed to be more reliable.

In their now classic statement of test theory, Frederick Lord and Melvin Novick (1968) are very clear about the limitations they see in the various educational and psychological assessments that had been developed in the previous part of this century. They argue that all tests which measure mental phenomena of real interest have little reliability, while all reliable tests measure trivial mental phenomena. Lord and Novick’s ideas are important beyond just a historical sense of the field of educational measurement because their position provided the support for later theorists like Lee Cronbach (1988) and Samuel Messick (1989), who would eventually revolutionize validity theory. In an important sense, then, “comparatively high reliability is neither a necessary nor a sufficient condition for establishing the validity of a measure” (Williamson 1994, 162). This distinction between the value of reliability and validity is crucial for those of us who work in writing assessment because of the dominant force reliability has been in the creation and design of writing assessment procedures. However, the importance of validity in educational measurement cannot be overstated, since the concept of validity furnishes the fundamental rationale for taking test results seriously and using them to make important educational decisions. In a nutshell, validity is the essential way in which a measurement is judged, what Pamela Moss (1998) has called “the test of a test.” The changing nature of validity reflects the changing values of those who work in education. Initially, validity meant whether or not a test measured what it purported to measure.

Traditionally, validity has been conceived as a three-part concept: content validity, criterion validity and construct validity. For a measurement to have content validity it must contain content and use procedures relevant to the phenomenon of interest. Essentially, the domain of the content being measured needs to correspond to the material used in the assessment. Criterion validity refers to the values and performances the assessment is designed to assess. For example, one type of criterion validity could be predictive validity, which means that a measure is able to predict how well a student will perform in some future context. SAT and ACT tests are said to have predictive validity, since they can reasonably predict student success in the first year of college. Writing assessment for placement into first-year writing courses would also need to have predictive ability because a reader examines student writing to make a judgment predicting where that student would most profit from
instruction. Other kinds of criterion validity include concurrent validity which stipulates a measure to be valid if it correlates well enough to another valued measure. For example, multiple choice tests of usage, grammar and mechanics achieved status as valid indirect measures of writing because scores on such tests correlated highly enough with scores given on holistically scored essays. Criterion validity can take many forms depending upon the assessment purpose and location. Within the trinity of content, criterion, and construct, construct validity was considered the most important of the three but also the most illusive. Early conceptions see construct validity as a net of relations that are connected nomologically (Cronbach and Meehl 1955) in consistent fashion to produce an adequate construct of the ability to be tested. More recent ideas about construct validity relate it to the conformity of a measure with a model or theory of the phenomenon of interest (Anastasi 1986).

It is important to note that validity does not reside within a test itself, but that “one validates not a test, but an interpretation of data arising from a specified procedure” (Cronbach 1971). The three-part nature of validity allowed claims about the validity of a particular assessment to be meted out piecemeal, so that something like a multiple choice test could be considered a valid measure of writing ability based upon the simple fact that scores on such tests correlate to a specific degree to scores received on holistically scored essays. This notion of validity as three separate entities often allowed validity to be conceived in rather limited terms. For example, Lorrie Shephard notes that during a thirty-year period from 1920–1950, “test-criterion correlations became the standard for judging test accuracy” (1993, 409). Writing in 1946, J.P. Guilford states that “a test is valid for anything with which it correlates” (1946, 410). Of course any of us who has had even an introductory course in statistics or rudimentary training in scientific methods understands the basic fallacy of confounding a correlative relationship with a causal one. This problem with reifying validity into its component parts was recognized in the 1974 version of Standards for Educational and Psychological Tests which is issued periodically by the American Psychological Association, the American Research Association and the National Council on Measurement in Education. These aspects of validity (content, criterion and construct) can be discussed independently, but only for convenience. They are interrelated operationally and logically; only rarely is one of them alone important in a particular situation (1974, 26).

While this made sense theoretically, in practice, separate camps within the measurement community continued to privilege one form of validity over another, allowing such practices as the validation of multiple choice tests for writing ability. In fact, writing as late as 1984, Peter Cooper draws upon this notion of validity in his statement about writing assessment: “From a psychometric point of view, it does appear that indirect assessment alone can afford a satisfactory measure of writing skills for ranking and selection purposes” (1984 27). By the 1980s Lee Cronbach (1988) and Samuel Messick (1989)
began to write a series of essays that changed the direction of validity study away from strictly technical and statistical considerations to a consideration of a range of social and environmental factors important in determining the validity of decisions made because of tests and in the formulation of a theory of validity as a singular concept. In this sense, validity can no longer be considered in three separate properties. Instead, in order for a test to be assumed valid it must meet standards for construct validity. Construct validity, then, subsumes matters of content and criterion, so that to speak of construct validity itself is to speak of validity. Lee Cronbach (1988), who has been influential in the evolution of validity theories since the 1950s, views test validation as the construction of an evaluation argument for the soundness of a particular set of inferences to be derived from a specific measurement. Validation speaks to a diverse and potentially critical audience; therefore, the argument must “link concepts, evidence, social and personal consequences, and values” (Cronbach 1988, 4). Cronbach’s emphasis on audience and argument lends a rhetorical feel to validity that resonates with those of us who work in writing assessment. Cronbach’s rhetoric of validity also resonates with current conceptions of language and literacy as he stresses the importance of the “social” and “personal,” recognizing that validity (as does rhetoric) contains “consequences” and values crucial to groups and individuals.

Cronbach’s insistence on linking the conceptual with the evidential foreshadows and mirrors Samuel Messick’s emphasis on the combination of empirical and theoretical concerns in test validation. In Messick’s definition, “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” (1989, 13). This definition of validity combines empirical and theoretical perspectives in constructing guidelines for making sound judgments based upon a particular measurement. In this sense, Messick is calling for a theorized practice of assessment in which theory and practice are inseparable from each other. This fusion of theory and practice is also central to the discipline of rhetoric and composition. James Zebroski (1994) articulates the centrality of the theory/practice connection:

Theory is not the opposite of practice; theory is not even a supplement to practice. Theory is practice, a practice of a particular kind, and practice is always theoretical. The question is not whether we have a theory of composition, that is, a view or better, a vision of our selves and our activity, but whether we are going to become conscious of our theory. (15)

In this sense, validity requires that we become conscious of our theory of assessment and of the theories that support the educational practices and decisions the assessment is designed to measure.

This intellectual move to become more conscious of the theoretical implications of our practices and the practical implications of our theories has
important consequences for the kinds of writing assessment procedures we build now and in the future. Pamela Moss notes that “Where the mainstream literature in validity theory (and program assessment) falls short, in my judgment, at least in practical emphasis if not in intent, is in encouraging a kind of critical reflection about its own taken-for-granted theories and practices” (1998 112). This critical reflection is even more crucial for writing assessment, since a full understanding of its issues and practices requires an interdisciplinary approach involving a multiple theoretical understanding. For example, in delineating a systems approach to validity, Fredericksen and Collins (1989) refer to work done in writing assessment to highlight the forward-thinking nature of composition and language arts teachers in linking educational outcomes and environment to assessment in writing. In general, this fusion of practice and theory can highlight the importance of pedagogy with educational measurement and promote the kind of dialectic that Moss (1998) and others like Donald Schön (1983) call for in conceiving appropriate professional activity as reflection-in-action. This discussion of multiple theories and practices in understanding writing assessment brings us back to an earlier point we mentioned about the importance of English teachers learning statistics or psychometrics to be able to carry on conversations with measurement specialists (Scharton 1996; White 1994). It should begin to be clear from our discussion of current theories of validity that English teachers’ understanding of the importance of rhetorical principles and the integration of theory and practice in composition positions them to carry on fruitful conversations with the measurement community. Although we don’t mean to ignore the statistical side of educational measurement, it is a misnomer to highlight statistics as the fundamental property in educational or writing assessment. Instead of emphasizing statistics, we would rather advocate a working knowledge of a few fundamental principles in validity theory for those English teachers who work in assessment, because emerging theories of validity have the potential to pressure measurement specialists to consider the importance of composition theory in writing assessment design and implementation.

COMPETING MODELS OF ASSESSMENT

Current traditional forms of writing assessment, like analytic, holistic, or primary trait, share common assumptions from traditional forms of measurement theory and practice. These procedures were developed by the CEEB and ETS during the first half of the twentieth century and were widely adopted during the 1970s. All three require that readers be trained to agree and to score papers according to a common rubric that describes numerical points. Each paper is given two different scores, and the scores of more than one scale point discrepancy are mediated to produce a more overall reliable rate of agreement. Such procedures share many of the assumptions of logical positivism, including the importance of technical and statistical rigor. The emphasis is on creating procedures in which readers can consistently score
the same paper. Rates of agreement are calculated and monitored because without a high enough rate of reliability, scores cannot be considered valid. Student writing is scored because it is assumed that student ability in writing can be communicated numerically. There are no restraints on the guidelines used to describe specific score points. Recent writing assessment programs like, the statewide assessment program in Georgia from 1989 to 1990, used a scoring guideline that was predominantly focused on language conventions (Englehard, Gordon and Gabrielson 1992). The scoring guideline for the assessment of state-mandated portfolios in the State of Kentucky uses the exact same score criteria for portfolios scored from students in elementary, middle, and high school grades (Kentucky Department of Education 1994), even though writing at these three different levels is likely to be quite different in terms of subject matter, skill, audience and purpose. Such consistency is in keeping with the purposes of current traditional writing assessment procedures like the trend sample of the writing portion of the National Assessment of Educational Progress, which attempts to measure and compare generations of students’ writing ability.

These traditional procedures, and the theoretical and epistemological rationales which highlight the importance of consistency and standardization, assume that these qualities can and should be maintained across temporal and environmental contexts. Traditional forms of writing assessment like holistic, analytic or primary trait are clearly supported by traditional notions of validity that focus on determining that a test measure what it purports to measure and that the consistency of a test be maintained at all costs. Lord and Novick’s (1968) concern from thirty years ago seems poignant in the consideration of the two examples of state wide tests from Georgia and Kentucky. Both testing systems provide an efficient means of examining a large number of students’ writing, and there is every reason to believe that this examination is consistent across grade and school district boundaries. What is not so clear, however, is how robust a version of writing ability is actually being examined. Can we assume in Messick’s (1989) terms that the large-scale assessment of student writing undertaken at the state or even national level provides the empirical and theoretical rationales for adequate and appropriate decision-making about student writing? Is there enough high-quality evidence to construct, in Cronbach’s (1988) terms, an evaluation argument for the validity of important decisions based upon these measures? Of course, a responsible answer to such questions would require research beyond the scope of an entire chapter in a single book. We do think, though, that the disjuncture between newer versions of validity and current traditional notions of writing assessment are clear and indisputable.

Recently, we in composition studies have begun to develop writing assessment procedures based upon our understanding of literacy and its teaching (Allen 1995; Durst, Roemer and Schultz 1994; Haswell and Wyche-Smith 1994; Smith 1993). Newer versions of writing assessment that do not share either the
assumptions of traditional measurement theory or the practices of current traditional writing assessment have been appearing in the literature for the last five years or so. William L. Smith developed one of the earliest of these new procedures for placement of students in the various courses in the composition program at the University of Pittsburgh. Smith’s design deviated from traditional procedures in that he hired his readers according to the classes they most recently taught. Instead of scoring student writing on a numerical scale, readers made one of two decisions, whether or not students should be placed in their course, or whether the student writing should be referred to the readers for the course above or below their own. This system involved no formal rubrics or training procedures, since readers made decisions about particular classes based upon knowledge they already had. Smith reported that this method allowed a more accurate and efficient placement of students than did holistic scoring, which had been used previously. Smith’s method has been adapted at the University of Louisville in a pilot program that uses students’ state-mandated portfolios for placement, and it has been found to work well with portfolio placement (Lowe and Huot 1997).

The placement program developed by Richard Haswell and Susan Wyche-Smith (1994) at Washington State University involves having only one reader examine all student placement essays. This reader decides whether or not a student would profit from the composition course designed to meet the needs of most enrolling first-year students. Sixty percent of all students are placed into the most heavily enrolled course based upon this singular reading. The other forty percent of students have their writing read a second time by a group of readers who were expert in all courses in the first-year composition curriculum. This multiple “tiered” reading, as it has been dubbed, has been extended to create a portfolio junior-level writing assessment program for all students (Haswell 1998). Again, these programs have proven to be more efficient and accurate than those they replaced which depended upon rubrics, training, interrater reliability, numerical scoring and the other trappings of what we call the technology of testing.

We now consider one last model in which students writing is read to make important decisions—this one from the University of Cincinnati. This program is designed to ensure that all students who move from one course to another have the necessary skills envisioned and demanded by their writing program. Competency programs, even those using portfolios, are not that rare these days. What makes this program worthy of mention is that student writing is read by three-teacher teams known as trios. Each teacher brings her students’ portfolios to be read together as a group. These trios make decisions about whether or not student writing is at a level commensurate with program standards. In addition to providing accurate decisions about students’ ability to move from one course to another, this program also provides important benefits to the program itself: “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 freshmen English program” (Durst, Roemer and Schultz 1994, 287). Such an assessment program meets the concerns of newer forms of validity that educational assessment have a positive influence on teaching and learning. Just as it would have been irresponsible of us to claim any sort of invalidity for more traditional writing assessment measures, it is equally problematic for us to claim any kind of validity for these newer writing assessment practices. However, we believe it is clear to see how these new programs attempt to combine the accurate generation of assessment data with an understanding of the ways in which student writing is read, evaluated and taught within acceptable educational contexts. It is conceivable, for example, that we could make an evaluation argument for some of these newer forms of writing assessment that links “concepts, evidence, social and personal consequences and values” (Cronbach 1988).

One implicit contention we hold about these new forms of assessment is that through the use of specific readers for specific educational decisions or the use of teacher collaboration to make decisions about program standards and values, what emerges is a type of writing assessment that links the making of decisions about written texts to specific contexts. The importance of context in meaning making through literate practices has its roots in various literary, linguistic, and rhetorical theories and theorists (for example, see Derrida; Halliday; Perelman). This theoretical grounding is also important to informed pedagogy in the teaching of writing that helps students define purposes and locate audiences for their own writing. As we have seen in our examples of newer forms of assessment, audience and purpose also play an important role in writing assessment. Our last example of writing assessment outside of a traditional theory of assessment also comes from outside an educational context and looks at the ways people in the publishing industry go about making important decisions about writing. Over the last decade or so Janice Radway has done a series of studies about the Book of the Month Club (BOMC). Much of Radway’s study has come from her focus in critical theory and literature about the way in which competing cultural values get inscribed and replicated in people’s reading habits as exhibited in the kinds of texts BOMC offers to its readership. What we are interested in, however, is the part of the study in which Radway describes how decisions are made about what books get chosen for the club itself. The assessment task the club faces is quite daunting: “its editors sift through more than five thousand manuscripts annually to choose approximately two hundred and fifty books that the club eventually offers to its membership” (1989, 260). What makes this task even more complicated is that the club offers selections in “several informally defined categories,” so that the selection process is not just about choosing the best books. Like the newer forms of placement we reviewed above, readers are choosing texts that “serve . . . different kinds of readers who use books for different purposes and thus judge them according to different criteria” (1989,
261). Just as traditional forms of writing assessment describe specific numerical points on a rubric, the BOMC isolates texts with specific criteria for its diverse readership. However, unlike traditional forms of writing assessment, the Book of the Month Club does not choose its books according to a set of articulated criteria, nor does it train its editors or stipulate that they code their decisions about texts into numeric rankings. Instead, editors read and select books, writing descriptions of their selections for the BOMC catalogue that is sent out to its readers on a regular basis.

With little more to go on than a paragraph or two of description and a few sketchy plans for future publicity, they must “place” the book, evaluate it, and decide whether they can convey their enthusiasm to a potential reader in a description that is routinely two hundred to three hundred words long and no more than one thousand. Their reading is governed finally not only by their own preferences and training but by the fact that the club is a business designed to sell books to others. They are always trying to read as they believe their members do. (1989, 262)

It’s interesting to note that these procedures resemble in some striking ways the newer forms of assessment we have been describing, even as they differ from traditional forms of assessment. The judgments of the editors for BOMC are based upon their experience and expertise as readers, as editors, as employees of a particular company, and as writers who must take their judgments about texts and convey them successfully to the readers who will ultimately purchase the books and eventually ratify the accuracy of such judgments. What makes the comparison of the way BOMC editors make decisions about texts so interesting to us is that while the similarities between new forms of educational writing assessment and differences from traditional forms are so striking, the situation and outcome of the decisions are so radically different. What appears to emerge here from a consideration of the assessment of texts from both an educational and professional context is a sort of professional ability to read texts for various purposes and to make important decisions without the need to articulate standards or ensure standardization.

These competing models of assessment, it seems to us, are based upon competing models of the way people arrive at judgments about written texts. Traditional measurement theory is based upon a Platonic universe and positivist epistemology in which there is an idealized universal truth that assumes a single correct answer. New forms of assessment assume the importance of context and purpose in the process of reading written texts, student or professional, to arrive at a specific judgment, whether that is the suitability of a book for a particular readership or whether or not a student should be placed into a specific class or exited from another. Regardless of the decisions in mind, there is an assumption that professional editors or teachers are qualified to make these types of decisions based upon their experience and expertise without the need of special training or monitoring. The ability to make
appropria\text{e}t deci\text{sions is in the expertise and experience of the readers, not in \text{the} external technology of a specific methodology or set of procedures.

\textbf{WHAT DIFFERENCES MAKE THE DIFFERENCES?}

Before we can answer the question that this concluding section of our essay asks, we need to articulate what the differences, indeed, are. The biggest difference, it appears to us, is the use of different terms which reflect different attitudes and assumptions about the way writing is assessed. Holistic, analytic and primary trait scoring are about what \textit{scores} students get or raters give. In newer writing assessment schemes teachers \textit{read} student writing to make decisions about writers. While traditional forms of writing assessment are based upon a \textit{theory of scoring}, new forms of writing assessment are based upon a \textit{theory of reading}. Scoring, as we have already discussed in our section about the theoretical basis for traditional writing assessment procedures, assumes a knowable, ideal, single account of writing quality that requires a specific technology like training, rubrics, etc. to make that assumption a reality. Reading, on the other hand assumes a divergent approach to meaning making, since different readers are capable of multiple accounts of a given text, depending upon the context of the reading and the position of the reader. Wolfgang Iser (1978) coined the term “wandering viewpoint” in his book \textit{The Act of Reading} to describe the flexible nature of a reader’s ability to comprehend various types of written discourse. Whether we consider literature from psycholinguistics like the work of Frank Smith (1982) or from reading theory like Richard Anderson and P. David Pearson (1984) or literary theory like Stanley Fish (1980), the current characterization of the act of reading is one of fluidity, adaptability and divergence. Each of the various newer methods for assessing student writing as well as the BOMC procedures are all based upon a notion of reading that privileges the individual reader who has the necessary expertise, experience and authority to make accurate judgments about written discourse. The key to the kind of reading given by an individual is the specific context within which she reads. In newer forms of writing assessment and the method used by the BOMC, the reader is given a rich context within which to read. Conversely, in traditional forms of writing assessment, raters are restricted to the limited context of the rubric and training they receive to agree with other scorers.

Edward Wolfe’s (1997) research into the scoring processes of highly reliable holistic raters indicates that they employ a more focused and limited reading process than less reliable raters. Inherent in the difference in context provided by the different methods for assessing student writing is the amount of control given to individual people or sites. There is a reluctance, we believe, on the part of traditional procedures to let such decisions about writing assessment be based upon such authority as an expert reader and his or her local community. Standardizing procedures and limiting context is another way to ensure control for assessment, standards and values with a central authority.
Radway notes that although BOMC editors did successfully choose books for the club, these decisions are not based upon extensive market research or surveys of readership taste or preferences.

The editorial decisions upon which the present study are based were still grounded predominantly in hunch, intuition, and luck. This is not to say that the editors knew little about their audience, only that the knowledge they had of it was not in quantifiable form but existed as the tacit, relatively unconscious product of long personal experience. (1989, 264)

Radway’s language here is revealing. In her use of such words as “hunch, intuition, luck,” she relegates these important, accurate assessments of texts to a sort of second class status because “the club’s past editorial operations have been based on surprisingly little hard data” (264). Her classification of these judgments parallel the reactions some have had about new forms of writing assessment.

A couple of years ago at the University of Louisville, we were contacted by an employee of the Kentucky State Department of Education who had heard we were using state-mandated portfolios to place students into first-year writing courses. She offered to train us according to the holistic rubric and methods the state uses for scoring the portfolios. We told her we were adapting William L. Smith’s method, and when she questioned the lack of a rubric we referred her to literature on the reading processes of holistic raters (Huot 1993; Pula and Huot 1993; Smith 1993) which seems to indicate that readers have an internalized rubric that they use to read and rate student writing regardless or in spite of any holistic training they might receive. Not too long afterwards, we heard through the office that acts as a liaison for the university to the public schools, that the State Department was concerned that we were letting teachers use their “intuition” to score writing portfolios for placement. Of course, we took great pains to correct what we saw as a misconception of our procedures. Our lack of comfort with characterizing our teachers as using their intuition points out the suspicion with which teachers’ decisions are often viewed and the lack of authority accorded to their expertise and experience. In fact, Edward White (1995) sees direct writing assessment as a way to control for the problems inherent in teachers’ readings of their students’ writing. In our specific case, we were aware that the design of our procedures was based upon research into the ways teachers read student writing in holistic scoring sessions; we were also aware of the ways in which standardized methods of assessing student writing are privileged over more local ones. What we also see at work here is that local, contextualized knowledge about the way people read and arrive at judgments about that reading is not considered to be as good or appropriate as procedures that are more standardized, that appear more scientific, objective or quantifiable. Even though we could argue for the validity of localized procedures within Cronbach’s ideas of validity as argument (1988), ultimately we have to argue against the continuing perception of superiority for standardized methods for
assessing writing. In other words, argument for newer forms of writing assessment are often not about the new methods but about ingrained perceptions of what assessment ought to be. Like Radway, we at Louisville were uncomfortable with characterizing important evaluative decisions about text as based upon intuition. While perceptions for assessment continue to rely upon more technological and scientifically based procedures, those of us promoting a new agenda for writing assessment need to recognize the power of an expert reader’s “intuition, hunch or luck.”

The critical importance of theoretical and epistemological differences between traditional and emergent writing assessment practices seems too large and important for a single essay. Clearly, the differences in praxis point out completely different theoretical and epistemological orientations, not to mention the important ideological question of where the power and control lie in these different versions of assessment. Our hope is that this discussion of theory and practice in writing assessment helps to bring to the surface the important task of looking past assessment practices to the theories that inform them. It is paramount, we believe, that those of us who teach student writers gain some control of the way student writing is assessed in and outside of the classroom. Calls for state and national assessment programs continue to be an important political rallying point for school-reform movements as well as for the advancement of individual political careers. Continuing the conversation about writing assessment which we have hoped to begin in this chapter is an important way for English teachers to have some say about educational reform and about the ways writing will not only be assessed but what about what kinds of written communication we will be able to value and teach in our classrooms. Conversations about assessment theory and practice need to continue and flourish, and these conversations need to include English teachers and writing program administrators. As we highlight the importance of theory and practice in composition studies as a whole, we also need to make a conscious effort to include such conversations about theory and practice in writing assessment.

NOTES

1. In her book *Institutionalizing Literacy*, Mary Trachsel (1992) examines the history of the essay examination in higher education in America and explores its impact on writing assessment and writing instruction.
2. These guidelines are published periodically to provide a set of professional principles for those working in educational assessment. These standards are currently under revision because of recent calls to make professional standards of practice in assessment relevant to more current theoretical understandings of validity and its subsequent effect on decisions and inferences based upon educational assessments.
3. Comparison studies of holistic, analytic and primary trait scoring show that all three procedures correlate fairly well (Freedman 1984; Veal and Hudson 1983). Holistic scoring has become the procedure of choice mainly because it is more efficient. (Veal and Hudson 1983). Actually primary trait was adapted from analytic scoring by a group working outside of the CEEB or ETS who were trying to implement an examination of writing for the National Assessment of Educational Progress. See Richard Lloyd-Jones (1977) for a description.

4. We are reminded of an experience we had at a meeting about writing not too long ago. After we described some procedures we had been using, one of the participants asked us what rubric we had been using. When we replied that weren’t using a rubric, his response was, “Oh, this is just a subjective assessment.” We, of course, pointed out to him and the rest of the audience that writing out criteria for assessing did not necessarily make an assessment more objective.