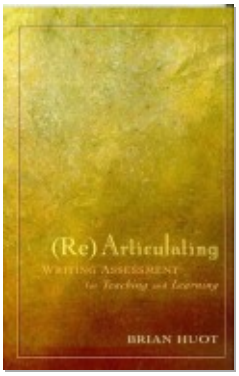
Rearticulating Writing Assessment for Teaching and Learning

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/.

For additional information about this book
https://muse.jhu.edu/book/9328

Access provided at 20 Sep 2019 09:31 GMT with no institutional affiliation

2

WRITING ASSESSMENT AS A FIELD OF STUDY

It is becoming more and more clear to me that the work that I and others do in writing assessment, like work in other fields, is constrained, shaped and promoted by the overall shape of the field itself, yet writing assessment—as a field—has not been the object of inquiry for very much scholarship. There are several reasons for this, of course. Writing assessment researchers have been busy doing other things, mainly trying to establish procedures that could measure student ability in writing. Although research into the assessment of writing goes back to the early part of this century (Starch and Eliott 1912), there really hasn't been much of a sense that writing assessment was, indeed, a field of study. Most work in assessment before the 1970s was carried out within the field of educational measurement, which still considers writing as just one more area of research within its vast domain of all educational testing. Interest and activity in writing assessment, however, has changed radically since the 1970s. In the last three decades, there has been much research and inquiry into writing assessment issues, enough by the early 1990s to support the establishment of a journal, *Assessing Writing,* devoted entirely to the assessment of student writing, and more recently a second periodical, *The Journal of Writing Assessment.*

Writing assessment has evolved into an intellectual and public site in which scholarship is conceived and implemented by people from various disciplines and subdisciplines. In a 1990 review,

I was able to identify three main foci for the existent literature in writing assessment: topic development and task selection, text and writing quality, and influences of rater judgment on writing quality (Huot 1990). I had attempted to let the issues covered in the literature itself focus the review and discussion, being very careful in choosing the three areas around which most scholarship in writing assessment clustered. Just four years later, in introducing the new journal *Assessing Writing* (Huot 1994c), I noted that none of the articles in our first issue dealt with the three major themes evident in the literature review four years earlier. The fact that areas of interest change in a given field of study is not by itself a significant point. However, what I noted then and bears repeating now is that scholarship in writing assessment up until the 1990s was mostly concerned with establishing the procedures themselves. While the landmark study conducted by Godshalk, Swineford and Coffman and published in 1966 outlined the procedures necessary to produce agreement among independent raters, the scholarship for the next twenty-five years or so focused on how to maintain the efficacy of these procedures (White 1994), as well as how to solve technical problems like the creation of similar topics for tests that attempted to compare scores from one year to another (Hoetker 1982), or how to train raters to agree and then statistically compute this agreement (Myers 1980). What's important to highlight in a chapter on writing assessment as an area of study is that while the literature up until the early 1990s focused on establishing and maintaining writing assessment procedures, more recent work has begun to critique current traditional writing assessment practices.

In her recent history of writing assessment, Kathleen Yancey (1999) notes that much writing assessment in the 1950s and 1960s was conducted through the use of multiple choice tests of grammar, usage and mechanics. Although essays had been used since the nineteenth century to test writing ability (Connors 1986; Traschel 1992), they had always been held suspect by the educational measurement community because of the low

consistency of agreement between independent raters—what is termed *interrater reliability*. As early as 1912, essay testing was proclaimed problematic because it was unreliable (Starch and Elliott 1912). It was not until 1941, however, under the pressure to test and matriculate students for World War II, that the College Board actually did away with essay testing (Fuess 1967). The establishment of the reliable procedures of holistic, primary trait, and analytic scoring for writing assessment in the 1960s and early 1970s was no small feat, and the attendant optimism it generated is understandable (Cooper 1977; White 1994). This optimism continued up into the 1990s, as most of the literature on writing assessment attempted to establish and maintain its legitimacy as a valid and reliable form of direct writing assessment. This reckoning of how the study of writing assessment is constituted is especially important for an area like writing assessment, since it is a subject that draws interest from a diverse group of people, from classroom instructors to writing program administrators, to school and universities officials, to state and federal legislators, to testing companies and assessment scholars, not to mention students and parents. Add to the conflicting interests of these groups the fact that work in writing assessment can come from different fields and various subfields with different and conflicting theoretical and epistemological orientations, and we get a picture of writing assessment as a field pulled in many directions by competing interests, methods and orientations.

TWO DIFFERENT DISCIPLINES

In understanding writing assessment as a field of study, perhaps the most significant issue is that many of the scholars involved represent different disciplines that hold differing and often conflicting epistemological and theoretical positions. Composition, of course, is a field that welcomes and uses knowledge from various fields and disciplines. However, in writing assessment, we not only borrow and use knowledge, but scholars from education and the measurement community consider writing assessment as their area of study as well. In fact, we are all in

debt to the measurement community for the most commonly used forms of writing assessment. Although Edward White (1993) and Kathleen Yancey (1999) write about developing or importing procedures from ETS, holistic scoring is essentially the same procedure *developed by* CEEB and ETS. It's important to note that while the reintroduction of essay scoring in the early 1970s was seen as a real breakthrough for composition and English teachers (White 1994, 1993; Yancey 1999), in reality it was the culmination of decades worth of research by the educational measurement community who had been grappling with the problem of reliability since the early part of the century. While I applaud the work that produced holistic and analytic scoring and the movement to use student writing in assessing student writing ability, it's important to see holistic scoring in two ways. The English teaching profession vociferously protested English and writing tests that contained no writing (Fuess, 1967; Palmer, 1960) and promoted continued research into essay scoring that culminated in the research (Godshalk, Swineford and Coffman, 1966) that produced acceptable rates of interrater reliability. However, no English or composition scholars played a major role in the development of holistic scoring. It was a procedure devised to ensure reliable scoring among independent readers, since reliability as a "necessary but insufficient condition for validity" (Cherry and Meyer 1993, 110) is a cornerstone of traditional measurement that spawned multiple choice tests and the entire testing culture and mentality that has become such an important part of current ideas about education. Although the advent of holistic scoring permitted student writing to once again be part of the tests in English and writing, we must not lose sight of the fact that holistic scoring is a product of the same thinking that produced the indirect tests of grammar, usage and mechanics. That is, like multiple choice tests, holistic scoring was developed to produce reliable scores.

Perhaps it's best to understand writing assessment as an area of study that is, at least in the ideal, interdisciplinary. I say ideal because interdisciplinarity involves an integration and dialectic

that has not been present in writing assessment study, though the type of borrowing across disciplines that sometimes occurs in writing assessment scholarship has at times been labeled interdisciplinary (Klein 1990). It might be best to call the scholarship in writing assessment *multi*disciplinary, since it has taken place within various disciplines. The idea of writing assessment existing across disciplinary boundaries is probably not new, though there is little crossover of scholars and work. For example, in 1990 I published two essays on writing assessment, one in *College Composition and Communication*, the flagship journal for composition, and the other in *Review of Educational Research*, a journal published by the American Educational Research Association, the major organization for educational researchers. Invariably, I have found that within a specific article either one or the other piece would be cited. Very few people ever referred to both, and of course it was easy once I saw which piece was cited to know what field of study the writer(s) represented. This example illustrates the lack of integration of scholarship within writing assessment. Writing assessment scholarship occurs in two academic forums, and the lack of connection between the two is a notion that we have yet to address in writing assessment literature because this literature has been written and read by those within a specific field who have little or no knowledge or interest in the other approach.

Edward White's essay, "Issues and Problems in Writing Assessment" (1994b) notes that people who work in writing assessment often have very different orientations toward testing and education. White's list of those with an interest in writing assessment: "writing teachers, researchers and theorists, testing firms and governmental bodies, students, minorities and other marginalized groups" underscores his point that the interests and approaches of the various factions in writing assessment put certain claims on what kind of assessments we should design and how these assessments should be used. This version of the field seems to be in line with the notion of the "stakeholder," assessment talk for all of those who have a claim on a specific assessment. In

White's view, each group competes with other groups for preeminence. While he gives some sense of the various tensions writing assessment needs to address, his categories dwell on the individual roles people play within the area of writing assessment and the needs and claims these roles suggest, without taking into account the larger social, disciplinary and historical factors that help to create the tensions he discusses, which in turn make writing assessment the field it is. It is fair to note that White's purposes in outlining the various people who may work in writing assessment and the various concerns that these people bring with them are different from my purposes here, since I'm interested in the makeup of the field itself. However, White's approach does have implications for the ways we think of the field. If writing assessment, as White suggests, is a field made up of various individuals who have differing and conflicting interests, then one implication is that we need to create a venue or forum that allows all of these concerns to be heard and addressed—which is exactly what White suggests at the conclusion of his article. This picture of the field and its suggestion for the future is not unlike the one depicted by Yancey's (1999) recent history in which she urges the balancing of reliability and validity as a way to reconcile disparate forces in writing assessment.

To understand the forces that both Yancey and White identify in their own ways, we need to look at the larger social, historical and disciplinary factors that comprise the field of writing assessment. Educational measurement is an area of study that can trace its roots back to the early decades of the twentieth century when researchers struggled not only to design and administer the first educational tests but also to establish the viability of the idea that there were, indeed, educational achievements and aptitudes that could actually be measured. This movement in educational measurement and its scholarship was closely allied to work in psychology which was also trying to establish the viability of certain human psychological traits that could also be defined and tested. The need to establish the viability and legitimacy of these enterprises and the fields themselves drove both educational and

psychological measurement scholars to consider emerging statistical procedures from the physical sciences. It's important to recall the intellectual climate of the early twentieth century and its focus on empirical, measurable, physical, human phenomena. Hence, the use of numerical explanations developed into the attendant field of psychometrics that attempts to understand human phenomena statistically. The connection between educational and psychological measurement can still be seen in the American Psychological Association (APA) *Standards for Testing*, which is published periodically and serves as a handbook for both educational and psychological testing and testers. It should be pointed out that although these standards are published by APA, the teams of scholars who write the standards come from both the educational and psychological testing communities. The field of educational psychology is an example of the interdisciplinary connections between these two fields. It should also be noted that scholars trained in educational psychology often work on questions regarding students literate practices and publish in journals devoted to English language education (Hilgers 1984, 1986; Shumacher and Nash, 1991), connecting educational psychology with the work done by education and composition scholars.

To understand the connection of writing assessment to the field of educational measurement, we probably should also understand its connection to the field of psychology. Recognizing these connections is crucial if we are to understand the theories behind current traditional writing assessment procedures like holistic, analytic, and primary trait scoring. Although I explore the nature of these theories more fully in chapter four as they relate to specific assessment practices, it's important to note that when Pamela Moss (1998) and others from educational measurement (Breland 1996) criticize college writing assessment, they are doing it from a theoretical perspective at odds with those I who work in composition. For if the educational measurement community is closely allied with the field of psychology, those who approach writing assessment from composition are allied with scholars in literary theory (Bakhtin 1981) critical theory

(Foucault 1977) and composition (Berlin 1988; Bizzell 1992; Faigley 1992). Just as important as the nature of these theories is the object of inquiry. For as educational measurement and psychology focus on sampling techniques, statistical trends and concepts like reliability and validity, composition looks at the importance of context and the processes of reading and writing and their teaching. It's safe to say, then, that White's categories of those who work in writing assessment contain individuals whose interests are shaped by certain theoretical and epistemological orientations and whose methods and approaches are determined by specific disciplinary allegiances. For example, when Yancey urges the balancing of reliability and validity, what she is really advocating is that those with different concerns for writing assessment, like English teachers and assessment specialists, work in harmony with each other. There is of course a certain logical, not to mention political, appeal to what White and Yancey see as the field of writing assessment and the collaboration they advocate. After all, if we are all stakeholders in writing assessment with our own competing claims, it is only by working together that we can honor these disparate claims.

The approaches to writing assessment advocated by educational measurement and college composition scholars are not only based upon different theories and epistemologies, but these approaches also value different aspects of an assessment. These different foci are recognized by large and reputable testing companies like the Educational Testing Service (ETS) and American College Testing (ACT) who regularly hire "content" staff who, for the purposes of writing assessment, have training and experience in literature, creative writing and/or the teaching of writing and language education. For the most part, content staffers in educational testing companies earn less than those trained in educational measurement, while those with backgrounds in educational or psychological measurement occupy supervisory and policy-making positions. In the world of professional writing assessment, then, there is some recognition of the need for information and expertise about the teaching of

writing, but that information is secondary to information about technical and statistical properties of developing, administering and interpreting writing assessments. This emphasis on the technical aspects of writing assessment visible in the structure of testing companies is also part of the literature on writing assessment (Breland 1996; Camp 1996; Scharton 1996) in which the technical aspects of writing assessment are emphasized, and English teachers' opinions about and efforts with writing assessment are criticized (Breland 1996; Scharton 1996). This criticism of college writing assessment by those with an interest and background in educational assessment signals that the work in college writing assessment by those with backgrounds in English and composition is at least starting to attract some attention by the educational measurement community, even if it is critical.

ISOLATION

Writing assessment has over the last three decades become a field in which scholarship takes place in different disciplines, and these two disciplines, college English and educational measurement, have different orientations, produce different kinds of assessments and are often in conflict about what constitutes appropriate writing assessment. The work of Pamela Moss, who is situated in educational measurement, has been used recently by people working in college writing assessment primarily because she has begun to challenge the status quo about reliability in her article "Can There be Validity Without Reliability" (1994a). More recently, in responding to Richard Haswell's scheme for validating program assessment, she notes that "Professor Haswell paints a picture of the field of college writing assessment that appears seriously isolated from the larger educational assessment community" (1998, 113). Like White, Moss sees writing assessment as an area of study in which different people pursue their own agenda, asking different questions and using different methods. Unlike White, however, Moss's notion of writing assessment is set in terms of disciplinary connections. Instead of pointing out various individuals who work

in assessment, her point about Haswell and others who work in college writing assessment is based upon their connections to a specific discipline. Haswell's work, Moss contends, could be made stronger were he to use approaches, principles and concepts from educational measurement. Moss's statement confirms the division in writing assessment between educational measurement and the college assessment communities. Moss's two categories for writing assessment correspond roughly to the divisions apparent in most testing companies in which content personnel familiar with writing and its teaching work alongside educational measurement specialists. I agree with Moss about the isolation of college writing assessment from the educational measurement community. In my own work, I have attempted to bridge the gap between educational measurement and composition because, like Moss, I see the value in much work done in educational measurement—not to mention the indebtedness of all of us who work in writing assessment to the research carried on for several decades that resulted in the development of direct writing assessment. Unlike Moss, however, I see the isolation she refers to as existing on both sides. So, while college writing assessment has been isolated from educational measurement, the converse is also true. Educational measurement has been isolated not only from college writing assessment but from the entire burgeoning field of composition.

To illustrate the isolation between educational assessment and college English, I look at work that attempts to outline writing assessment history, since historical inquiry can be a powerful indicator of disciplinary allegiance and because, like this chapter, histories tend to account for how a field came to be. In this sense, we look to the past to understand why certain ideas, principles, practices, theories and people are important to our present work. Looking at the way certain scholars configure writing assessment history is an indication not only of the values they hold as a individual members of a specific community but of the values of the community itself. Writing assessment is a complex historical subject because it is an area that remains multidisciplinary, drawing

scholars and interest from across disciplines and fields. For my purposes, histories of writing assessment provide an interesting picture of the kinds of isolation or multidisciplinary interaction that create the current climate of the field. The next section, then, explores three different historical views of writing assessment. In these, I look for points of isolation and intersection not just to get a sense of the field as it now stands, but to be able to make some suggestions for the future state of the field with less isolation, more collaboration, and better assessments for teachers, students and all of those affected by writing assessment and its influence on teaching and learning.

Edward White has been the preeminent college composition scholar in writing assessment for three decades. His book, *Teaching and Assessing Writing*, published first in 1985 and in a second edition in 1994 (the edition I refer to throughout this volume) is easily the most popular source for information about writing assessment for college-level writing teachers and program administrators. Although White has never published a history of writing assessment, his retrospective essay "Holistic Scoring: Past Triumphs, Future Challenges" (1993) serves our purposes here by providing a description of the role holistic scoring and writing assessment have had on college-level writing instruction and program administration during the decades of the 1970s and 1980s.

White frames the early work he and others in the college writing community accomplished in the 1970s as a "missionary activity" (79). This missionary activity was in response to the prevalence of multiple choice tests for the measuring of writing ability at that time. According to White, he and others were involved in combat with ETS officials and college administrators to find more accurate and fair ways to assess student writing ability. Holistic scoring fit the needs of White and others, since it was a method for scoring student writing that "could come under the control of teachers" (79). White outlines the differences he sees in the way that holistic scoring sessions are run by testing companies like ETS and those that he and other English teachers

administer: "It [ETS] tends to see too much debate about scores (or anything else) as time taken away from production of scores. But as campus and other faculty-run holistic scorings became more and more common, the warmth and fellowship they generated became one of their most valuable features" (88). Not only was holistic scoring a strong response to multiple choice tests, it also provided an important model for writing teachers themselves, since White claims that "When these same writing teachers returned to their classrooms [after holistic scoring sessions], they found that their teaching had changed. . . . [T]hey were able to use evaluation as a part of teaching, a great change from the customary empty whining about their responsibility for grading and testing" (89). The benefits of holistic scoring for teachers go beyond an attitude change toward assessment and provide them with models for assignment construction, fair grading practices, and the articulation of clear course goals.

White defines validity as "honesty and accuracy, with a demonstrated connection between what a test proclaims it is measuring and what it in fact measures" (90). He goes on to claim that holistic scoring is more valid than indirect measures. Most of White's discussion of validity has to do with ways in which holistic scoring might be less valid through shoddy task and prompt development and the inappropriate use of other testing procedures. His emphasis is on the technical features of the assessment itself. White's treatment of reliability is much more extensive than his discussion of validity; he devotes a little more than three pages to validity and almost seven to reliability. He acknowledges that "Reliability has been the underlying problem for holistic scoring since its origins" (93). White's conception of reliability, as I discuss in chapter four, is equated with fairness: "Reliability is a technical way of talking about simple fairness, and if we are not interested in fairness, we have no business giving tests or using test results" (93).

White's treatment of reliability underscores his belief that writing assessment is a site of contest and struggle. Ultimately, White's own position about reliability is ambiguous, as he ends

his essay with a discussion of portfolios and reliability: "While reliability should not become the obsession for portfolio evaluation that it became for essay testing, portfolios cannot become a serious means of measurement without demonstrable reliability" (105). The contradictory impulses in White's essay are part of what I take as a love/hate relationship not only within White's notion of reliability, but evident even in the way that college English views writing assessment and the researchers who developed methods like holistic scoring. On the one hand, holistic scoring is seen as a powerful technique with the capability to effect "a minor revolution in a profession's approach to writing measurement and writing instruction" (79), while on the other hand, ETS is seen as a powerful force which must be resisted: "To this day, some of the ETS people involved do not understand why the community of writing teachers and writing researchers were—and are—so opposed to their socially and linguistically naïve work" (84). Interestingly enough, White refers only to ETS, ignoring any part that the educational measurement community might have had in developing direct writing assessment measures: "Aside from one book published by the College Board (Godshalk, Swineford & Coffman 1966) and a series of in-house documents at the Educational Testing Service, I found only material of questionable use and relevance in statistics and education" (81).

Kathleen Yancey's (1999) essay "Looking Back as We Look Forward: Historicizing Writing Assessment" appears in a commemorative issue of *College Composition and Communication* (*CCC*), celebrating the fiftieth volume of the journal, the main publication of the Conference on College Composition and Communication. Yancey's and White's notions of the field of writing assessment are through their understanding of composition studies. White begins his history of writing assessment in the early 1970s, around the time holistic scoring became an assessment option and around the same time rhetoric and composition began to come together as a field. Yancey begins her history in 1949, to coincide with the initial publication of *CCC*.

While White's narrative begins in the 1970s with English teachers awakening to the realities, challenges, and potential of getting involved in writing assessment. Yancey acknowledges assessment as an important but invisible part of writing instruction in the 1950s. For Yancey, writing assessment from 1950 onward can be seen as three consecutive waves: from 1950–1970, objective tests; from 1970–1986, holistically-scored essays; from 1986–present, portfolio assessment and programmatic assessment (483). Yancey sees the wave metaphor as a way to "historicize" the different "trends" in writing assessment: "with one wave feeding into another but without completely displacing the waves that came before" (483). Like White, Yancey sees the history of writing assessment as a struggle between teachers and testers: "the last fifty years of writing assessment can be narrativized as the teacher-layperson (often successfully) challenging the (psychometric) expert" (484).

Like White, Yancey sees the early layperson assessment pioneers as having a major role in the development of writing assessment procedures:

> Which is exactly what White and others—Richard Lloyd-Jones, Karen Greenberg, Lee Odell and Charles Cooper, to name a few—set out to do: devise a *writing* test that could meet the standard stipulated by the testing experts. . . . Administrators like White thus borrowed from the Advanced Placement Program at ETS their now familiar "testing technology." Called holistic writing assessment, the AP assessment, unlike the ETS-driven placement tests, was a classroom-implemented curriculum culminating in a final essay test that met adequate psychometric reliability standards. . . . By importing these procedures, test-makers like White could determine both what acceptable reliability for an essay test should be and, perhaps more important, how to get it. The AP testing technology, then, marks the second wave of writing assessment by making a more valid, classroom-like writing assessment possible. (490)

This version of how English teachers came to control and use holistic scoring is remarkably like the one offered by White.

One main difference is that White notes that ETS scoring sessions are fixated on the delivery of reliable scores whereas those run by him and other English teachers permit a more convivial and community-building atmosphere. Yancey, on the other hand, sees the main difference in that English teachers borrowed AP testing that is more closely allied to a specific curriculum, since AP testing (which ETS continues to conduct) is designed to measure how well high school students have mastered a specific course of study.

Another similarity between White's and Yancey's version of writing assessment history is that they both claim validity for direct writing assessment, and that their versions of validity "Validity means you are measuring what you intend to measure" (Yancey 487) are pretty much the same. However, instead of the contradictory impulses we see in White's attitude toward reliability, Yancey, notes that "Writing assessment is commonly understood as an exercise in balancing the twin concepts of validity and reliability" (487). Yancey goes on to suggest how the various waves she defines have affected the relationships between validity and reliability. In the first wave of "objective" tests, reliability was the main focus. In the second wave, validity became the focus. In the third wave, validity is increased because "if one text increases the validity of a test, how much more so two or three texts?" (491). She is careful to note the continuing role of reliability. Using Elbow and Belanoff's early work with portfolios at SUNY Stony Brook as a model, she contends that "psychometric reliability isn't entirely ignored" (492) as readers "are guided rather than directed by anchor papers and scoring guidelines" (493). In this way, I see Yancey as trying to talk about writing assessment in ways that are more amenable to the educational measurement community, since she attempts to characterize developments in college writing assessment in terms like reliability and validity that have their origins in, and continue to have important meaning for, the educational assessment community.

Despite Yancey's attempt, the picture both she and White paint of college writing assessment conforms to Moss's point

about the mutual isolation between college writing assessment and educational measurement communities. In both accounts, English teachers are hero combatants who wrestle away control for writing assessment from testing companies who would ignore the need for writing assessment even to include any student writing. Whether we listen to White's version which has us see the faculty-run holistic scoring session as a virtual panacea for creating community, educating writing teachers, and producing accurate and fair scores for student writing, or to Yancey's, which distinguishes between "ETS-driven placement tests" (490) and holistic scoring developed for AP testing, the procedures under discussion were developed by the educational measurement community. And these procedures were basically identical, though I agree with White about the difference between an ETS scoring session and one run by English faculty. These test developers, like Fred Godshalk who coined the term "holistic scoring" in the early sixties, mostly worked for ETS. These are the people who experimented with training readers on scoring rubrics, so that independent readers would agree on scores at a rate that was psychometrically viable. This research culminated in the landmark study, conducted by Godshalk, Swineford and Coffman (1966) and published as a research bulletin by ETS in 1966, in which independent readers were finally able to score student writing at an acceptable rate of reliability. The procedures used in this study, like rater training on numerical scoring rubrics became the technology of direct writing assessment that continues to be used today. The educational measurement community created direct writing assessment as they had created the indirect tests. While pioneers in writing assessment outside the educational measurement community like Edward White (1994), Charles Cooper (1977) and Richard Lloyd-Jones (1977) struggled to implement the new procedures in a variety of situations and to bring them under English faculty control, the procedures themselves were created by educational measurement specialists working for ETS to provide a reliable way to score student writing.

Unlike White (1994), Yancey (1999) attempts to portray developments in writing assessment through the concepts of validity and reliability, with reliability being the main focus of indirect tests, and validity being the focus of direct writing assessments, like holistic scoring. Yancey contends that validity "dominated the second wave of writing assessment" (489) which Yancey pinpoints as the holistically scored essay during the time period between 1970 and 1986. However, Yancey's contention is not supported by the literature on writing assessment. While working on my dissertation in the fall of 1986, I conducted a complete Educational Research Information Clearinghouse (ERIC) search and found 156 listings for writing assessment in the entire database. Of these, over sixty percent were devoted to reliability. It was clear to me then, and it's clear to me now, that reliability dominated scholarship on writing assessment during that time period. I can think of two reasons why Yancey mistakes the 1970–1986 time period in writing assessment as being dominated by concerns for validity. One might be that she, like White, sees holistic scoring and other direct writing assessment procedures as a victory (coded as validity) won by English teachers over the educational measurement community (coded as reliability). Therefore in her mind, the proliferation of holistic scoring[1] allowed validity to dominate. A second reason might be that, because of her disciplinary affiliation, she is isolated from the scholarship on reliability and its connection to holistic scoring and other direct writing assessment procedures. The struggle that resulted in the development of holistic scoring took place within the field of educational measurement, since both indirect and traditional direct writing assessment were developed and designed to address problems in reliability caused by independently scored essays.

One of the biggest points of isolation between college writing assessment and educational measurement is in the treatment of validity. Both White and Yancey posit the outmoded definition of validity as a test that measures what it purports to measure. This impoverished definition allows for claims of validity regardless of

the theoretical orientation of the assessment or its consequences. For example, the recent writing assessment used by City University of New York can be pronounced valid, since the consequences of denying university entrance to scores of minority students does not interfere with what the test purports to measure. The test continues to be used to deny educational opportunities to students even though there is a body of evidence that shows that students who worked in developmental and mainstream programs were able to pass "the core courses at a rate that was even higher than the rate for our pilot course students who had placed into English 110" (Gleason 2000, 568). As I note throughout the volume and expand later in this discussion, validity has for decades meant more than whether an assessment measures what it purports to measure. Currently, validity focuses on the adequacy of the theoretical and empirical evidence to construct an argument for making decisions based upon a specific assessment. In contrast to the picture of validity offered by White and Yancey, over thirty years ago the educational community had already established an alternative concept of validity: "One validates not a test but the interpretation of data arising from a specific procedure" (Cronbach 1971, 447). White and Yancey assume the validity of direct writing assessment, with Yancey attributing increases in validity to assessments that are more "classroom-like" (490) or that contain multiple texts (491). Unfortunately, the validity of holistic scoring, the most popular form of direct writing assessment, has been asserted but never established—a point made by Davida Charney (1984) nearly two decades ago and seconded by me a few years later (Huot 1990). Whether or not validity as a guiding principle for assessment is something writing assessment should pursue is a separate issue and one I address later in this chapter. Nonetheless, it is clear from the two pieces by White and Yancey that college writing assessment has held a very different version of validity from that currently advocated by those in educational measurement.

Our somewhat cursory examination of scholarship from college writing assessment reveals, as Moss indicates, its isolation

from educational measurement. However, the converse is also true. Work on writing assessment from educational measurement exhibits an isolation from college writing assessment. To illustrate the isolation of educational measurement from college writing assessment, I choose to look at Roberta Camp's (1993) essay, "Changing the Model for the Direct Assessment of Writing" which was first published in an anthology I edited with Michael Williamson; a briefer version was published three years later in an anthology edited by Edward White, William Lutz and Sandra Kamuskiri. Camp is a well-known figure in writing assessment, working for ETS through most of her professional career. Her early work on portfolios was influential in making them such a popular writing assessment option. This essay, while not strictly a history of writing assessment, suits our purposes well as Camp outlines in the beginning, "The discussion will begin with a reflection on the history of writing assessment in recent decades and then go on to examine the current status of existing models for writing assessment" (46).

Like Yancey, Camp sees writing assessment history as balancing the requirements of reliability and validity. She explains how multiple choice tests of writing ability measure writing and how these tests claim validity.

> The multiple choice test, with its machine-scoreable items, provides evidence taken from multiple data points representing relatively discrete components of the writing task each measured separately . . . The claims for its validity have rested on its coverage of skills necessary to writing and on correlations between test scores and course grades—or more recently between test scores and performance on samples of writing, including writing generated under classroom conditions. (47)

Camp notes that these claims for the validity of multiple choice tests of writing were more persuasive "to statistically oriented members of the measurement community than to teachers of writing" (47). While Camp is ultimately sympathetic to those who would question the use of multiple choice tests, she does

note that there is some foundation for the validity claims of indirect measures of writing. Camp explains that eventually research indicated that although student scores on multiple choice tests and essay exams would be similar, that these "formats," as Camp calls them, were ultimately measuring different "skills."

Although Camp is a proponent of direct writing assessment, she is guarded in her claims for its validity: "In many respects, the holistically scored writing sample fares better than the multiple choice test with respect to validity . . . It has therefore been seen by some writing assessment practitioners as a stand-alone format for more valid assessment, especially when more than one writing sample is used" (49). For Camp, "the estimated test reliability for a single essay scored twice is insufficient to fully justify the use of a single essay as the sole basis for important judgments about students' academic careers" (49). The solution to this problem with the reliability of holistically scored essays is, in Camp's terms, a "compromise" which entails using both multiple choice tests and holistically scored essays. Camp acknowledges the importance of direct writing assessment and the many advances that have been made in the ways "we conduct evaluation sessions and report the results" (51).

In reflecting upon the history of writing assessment, Camp also attempts to look at the assumptions behind the procedures. She contends that many of the procedures designed to make writing assessment reliable might contribute to a questioning of its validity, since the streamlined process of having students write to identical prompts in test-like conditions only represents a portion of what we consider to be the skill of writing. This is in contrast to her assertion that a single-scored essay lacks the reliability to be valid. Camp refers to literature about the complexities of reading and writing that have emerged in recent years and concludes that both multiple choice tests and impromptu essays are lacking in their ability to measure the complexity involved in writing: "Neither the multiple-choice test nor the impromptu writing sample provides a basis in the assessment for obtaining information about the metacognitive aspects

of writing, information that is essential to instruction and the writer's development" (58). Camp contends that the more traditional forms of writing assessment are inadequate given the many recent breakthroughs in research, theory and practice about written communication: "The multiple choice test and the writing sample seem clearly insufficient for measuring writing ability as we now understand it" (58). For Camp, advances in knowledge about reading and writing have fueled advances in the assessment of student writing. Camp advocates that "we need to develop a conceptual framework for writing assessment that reflects our current understanding of writing" (59).

Camp contends that "the recent developments in cognitive psychology that have stimulated new perspectives on writing have brought new views of intellectual behavior and learning to all of education, including the field of assessment" (60). She focuses on changes in validity which no longer rest on the coverage of an assessment and comparisons to performances by students with other measures—the methodology implemented to justify the use of multiple-choice tests for measuring student writing. Instead, Camp asserts "that all evidence for validity is to be interpreted in relation to the theoretical construct, the purpose for the assessment, and therefore the inferences derived from it, and the social consequences" (61). The question for validity is no longer just whether or not a test measures what it purports to measure but rather "whether our assessments adequately represent writing as we understand it" (61).

Camp urges the creation of new models for writing assessment that capitalize on the continuing development of more complex understandings of literacy and its teaching. Combining a theory of learning which is emerging from cognitive psychology with recent developments in validity should allow us to create assessments for writing "that lead far beyond the narrow focus on score reliability and the constricted definitions of validity that characterized earlier discussions of the measurement properties of writing assessments" (68). Focusing on research about the composing process and building upon the lessons we

have learned through the use of direct writing assessment should provide a productive future for writing assessment and the creation of new, alternative models. She outlines three stages for the development of these models that focus on, first, identifying the competencies to be assessed and specifications for "the tasks to be presented" (69), second, "exploring scoring systems, and further refining tasks and scoring systems" (69), and third, "training readers" (69), "scoring samples," (69) and "conducting statistical and qualitative analyses to establish reliability, validity and generalizability" (69). Camp contends that "Procedures such as these suggest an orderly and responsible approach to developing and trying out new assessments of writing" (69).

Camp ends her chapter by pointing out that writing assessments are often lauded in educational measurement circles as being "exemplary as models for assessment" (70), since they attempt to represent and model the complexities of literate behavior. In characterizing what she sees as the future of writing assessment, Camp forecasts several features that have to do with creating an increased context within which student writers can work and in providing assessment activities and results that are more meaningful to teachers' professional development and understanding of their students' abilities. She also notes an increased attention and awareness of the cognitive processes involved in writing.

If, in college writing assessment history, English teachers are combatants in a struggle to wrest away control of writing assessment from testing experts, they are non-players in the historical accounting from educational measurement. They might have concern for including writing in its assessment (47), and, like White, Camp thinks that "No responsible educator would want to see a return to evaluations of writing based on the private idiosyncrasies of the individual evaluator" (58), but in Camp's history they otherwise have no role in writing assessment. Ignoring the role of early college writing assessment pioneers in this way not only dismisses their contributions but it also misses the development of a culture and advocacy that would eventually clamor for

assessments more compatible with writing instruction, and that eventually lead to portfolios (Elbow and Belanoff 1986) and writing assessments that go beyond the psychometric paradigm (Allen 1995; Haswell and Wyche-Smith 1994; Smith 1993).

For Camp, writing assessment has consisted of a "compromise" between multiple-choice tests and holistically scored essays. Camp even admits that holistically scored essays neither represent the complexity of writing nor do they, by themselves, satisfy the measurement requirements for reliability. Given this description of holistic scoring, I have to wonder exactly who or what is being compromised? We have multiple choice tests of usage and grammar that involve no writing or reading at all (though Camp contends that they do sample relevant content-area knowledge), and holistic scoring that according to Camp under-represents the process of writing while at the same failing to achieve necessary reliability. Unfortunately, I and others in college writing assessment would see no compromise here, but rather a continuing, unrelenting march toward reliability at the expense of validity—and complete dismissal of those outside educational measurement.

A continuing theme throughout this chapter is Camps's assertion that writing assessments began to change as our understanding of the complexity of writing became more apparent. This is a progressive agenda for writing assessment development that is driven by the knowledge we have about writing itself. It is also a view in which the responsibility for the problematic assessments of the past rest with content-area professionals, since once content-area professionals began to supply a more accurate and complex picture of the act of writing, assessments were developed to match. However, Camp offers no evidence for this assertion; she merely correlates advancements in writing assessment with those in literacy studies. Her position ignores the theoretical entrenchment of many in the educational measurement community. And, as I argue in chapter four, it is the beliefs and assumptions behind theoretical and epistemological positions that drive writing assessment practice. For example, as late as

1984, her colleague at ETS, Peter Cooper, writes, "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" (27). Keep in mind that this publication date is after many of the landmark studies on the writing process which Camp cites as being influential in promoting the development of new writing assessments. Even as late as 1998, Roger Cherry and Steven Witte write about the under-representation of writing in most assessments. Camp's assertion about the preeminent role of content knowledge about writing and literacy is an interesting and important idea that I hope guides writing assessment in the future, since it positions content-area knowledge in a leadership role. Currently, however, content-area professionals in testing corporations play a subordinate role; theories of testing—and not of language—drive most current writing assessments (see chapter four for a discussion of the theories that drive current traditional writing assessment).

Ironically, Camp does not mention assessments developed upon theories of language rather than testing (Allen 1995; Haswell and Wyche-Smith 1994; Lowe and Huot 1997; Smith 1993). While this work appears in print after or concurrent with the publication of this essay in 1993, Camp's second version of this essay in 1996 contains no references to this work. For me, Camp's neglect of work in assessment that calls for the very principles she advocates is due to her isolation from the college writing assessment community. This isolation can also be seen in the absence in her discussion of the influence of composition as a burgeoning field during the last three decades in which the direct assessment of writing has been evolving. Not only does Camp's isolation prevent her from tapping into new developments in writing assessment, but it also causes her to miss much of the new emphasis in language and literacy studies on the social nature of literate behavior (Berlin 1988; Bizzell 1992; Faigley 1992; and many others). Instead, Camp refers repeatedly to the advances in cognitive studies about the complexities of the way students write and learn. Her isolation from

college writing assessment, then, causes her to miss the development of new, language-based writing assessments and the continuing appreciation of the social aspects of literacy and its teaching, which have been a continuing focus in the composition literature since the mid-1980s.

This review is necessarily brief and incomplete, but it nonetheless shows how college English writing assessment professionals and educational measurement professionals—the two major communities responsible for the ongoing development of writing assessment—have been isolated from each other. Neither community has given the other the credit or respect for its accomplishments and contributions. We fail to recognize the debts we have to each other or the ways in which work in one area is stunted by its isolation from the other. While English teachers who work in assessment have often portrayed researchers in educational measurement as the bad guys (Elbow 1996; White 1994), more recent work from educational measurement refers to the efforts of English teachers as unreasonable and naive (Breland 1996; Scharton 1996).

VALIDITY

I began this chapter by noting that the work being done in writing assessment is constrained, shaped and promoted by the overall shape of the field itself. Tracing the two main influences on writing assessment, it is easy to see how both college writing assessment and educational measurement have been the prime shapers of what we know as writing assessment theory, research and practice. My examination of the way writing assessment history is represented by those in college writing assessment and educational measurement reveals two different versions of the field itself. As my discussion of the two historical representations indicates, both sides of the assessment coin are partial and limited; neither provides a complete enough picture of the complexities, issues and resources necessary to move writing assessment forward as a field of study. The isolation that Moss (1998) notes not only hinders any work undertaken in either

side of the field but, I believe, limits the field itself. I am not advocating a quick and dirty effort, by which writing assessment simply combines approaches from the two fields; there are conflicts and tensions that hardly fit together. White (1994) and others have already suggested a stakeholder approach that attempts to address the disparate concerns of people working in different fields for the assessment of student writing. What we need are some new directions for writing assessment scholarship, and practices that involve and attract both factions interested and invested in writing assessment. I hope to provide a framework that will make it possible for all those who work in writing to create new ways of theorizing and practicing assessment. While I outline this framework here, I write also in chapters four, six, and seven about the importance of validity theory in directing writing assessment toward a stronger role in promoting teaching and learning.

I begin this discussion of how we might create a new framework for writing assessment with a discussion of validity. Both sides of the writing assessment community talk about validity, though they talk about it different ways. The first step, then, is to take a closer look at validity. Although I see validity as an important part of all writing assessment, it cannot by itself mend the isolation in the field or provide a productive future for writing assessment. But it can perhaps provide a unifying focus that permits those in different fields to bridge gaps and make connections. It is possible that validity can be a way to make all those who work in writing assessment responsible to a given set of principles. Of course, it might be said that this has always been validity's role and that given the discussion of the field so far, it has failed, since we cannot at this juncture in writing assessment even agree upon what validity is, let alone agree to abide by its principles. There are two principles, however, that might enable us to use validity as a linchpin for holding together writing assessment, preventing the current isolation of those who now work in the area and charting a productive future. First of all, we need to agree on what validity is—to

decide what the principles for working in writing assessment should be. Secondly, we must hold each other responsible for following these principles. Although validity theory has been developed within the educational measurement community, that community has not always worked within the theoretical framework it provides (Shepard, 1993).

Both White (1994) and Yancey (1999) define validity as making sure a test measures what it purports to measure, though neither of them cite any source for this. This textbook definition can be found in many discussions about validity, but such a definition by itself is inadequate for many reasons. As Yancey notes in using the definition and citing F. Alan Hanson's book (1993) *Testing Testing: The Social Consequences of an Examined Life*, an assessment can create the categories for which it assesses. For example, I could use holistic scoring to make decisions about which students at the University of Louisville, where I work, have adequate writing skills to exit the first-year composition sequence. We could claim that the test measures what it purports to measure, since it would involve students writing and teachers reading that writing. We could cite adequate levels of interrater reliability, a scoring rubric that is general enough not to evoke any argument about its descriptions of writing ability, and the other trappings associated with holistic scoring, and eventually that test would come to represent writing ability in the first-year composition sequence at the University of Louisville. Instructors would begin including instruction on passing the test as a part of their curricula, so that their students would be successful on the test and even more importantly would be deemed good writers. Eventually, this test of writing would be the marker of good writing, de facto a valid test. All of this would take place without any attention to the decisions being made on the basis of this assessment. The assessment would exist outside of any determination about its impact on the writing, education, or lives of the students required to take the test, not to mention the test's impact on the curriculum of the course that students take before being tested. This example

points out that such a limited definition of validity is not only inadequate, it is dangerous because it accords an unexamined authority to an assessment that has the power to define educational achievement and influence instruction. My make-believe scenario closely resembles the real-life example of the ways in which the CUNY placement test is used to deny entry to many students (Gleason, 2000), underscoring the problematic nature of any form of validity that does not consider the consequences of the decisions made on the results of an assessment.

Since the 1950s, validity has been defined in more complex and comprehensive ways that attempt to provide more and more information not only about the test itself, but also about the theoretical framework that supports specific testing practices and the consequences on students and schools that result from the decisions made on the basis of the test. Before the 1950s, a more simple and reductionist notion of validity prevailed. For example, in 1946, J. P. Guilford states that "a test is valid for anything with which it correlates" (429). This notion of validity resting on a correlation to an outside criterion eventually became known as criterion validity. As validity theory developed, criterion became one of three forms of validity about which Robert Guion (1980) coined the term "the holy trinity." Criterion validity refers to the relationship of a measure to outside and relevant criteria. The second form of validity was content validity which pertained to the domain of knowledge, ability, or trait being measured. The third form of validity was called construct validity, and referred to the construct of the ability, skill, or performance being measured. For example, in writing assessment, the question would be whether or not the assessment contained an adequate construct of writing ability.

While certainly more complex than earlier definitions of validity, the trinitarian notion of validity had other shortcomings. Although content, criterion, and construct validity were never meant to function independently of each other, they were often reified and used independently, so that test developers could assert validity for a measure even if it were only a partial

claim based upon content or criterion validity. For example, as Camp (1993) details, multiple choice tests of grammar and mechanics based their claim for validity on both content and criterion validity. It was asserted that testing students on grammar and mechanics (later indirect tests even added questions about the writing process or rhetorical decision-making) sampled relevant content-area knowledge. Claims based on criterion validity noted that student scores on multiple choice tests correlated to some extent with the scores these same students received on essays they had written. Consequently, validity was asserted for multiple choice testing of an ability (writing) without there being any writing in the assessment itself. Of course, had the test developers considered construct validity, no such claim would have been possible.

Eventually, measurement scholars and validity theorists proposed a unified version of validity under the construct validity framework within which considerations of content and criterion would be subsumed. The intent of formulating validity as a unitary concept was to prohibit the parceling out of validity piecemeal to allow partial claims for the validity of an assessment. In other words, even though a multiple choice test could claim that it sampled relevant content from the writing process or that scores on the test had certain levels of correlation with scores on essay exams, a claim for validity would have to contain evidence that the exam represented a viable construct of the act of writing—a most difficult claim for a writing test that contains no writing.

Validity as a concept, then, has evolved from a simple correlation to "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" (Messick 1989a, 5). Messick's definition is widely cited and accepted in the educational measurement community. I use it, then, as a way to understand how his definition of validity contrasts with the one currently used by the college writing assessment community and to emphasize how Messick's notion of validity can be applied to

writing assessment. One of the biggest differences between Messick's version of validity and that commonly used in college writing assessment is the amount of information and activity necessary to satisfy a claim for validity. Instead of just looking at whether or not an assessment measures what it's supposed to, Messick's definition requires that we collect empirical evidence and furnish a theoretical rationale. Validity centers not on the measurement itself but on the "adequacy" of the decisions and "actions" that are taken based on the assessment. In this way, validity cannot be seen as a singular blanket that covers any assessment procedure like holistic scoring or portfolios. Information about decisions to be made and actions to be taken need to be supplied for each use of the assessment, negating not only a simple declaration of validity for a specific type of assessment, but introducing the necessity of supplying empirical and theoretical evidence of validity for specific environments, populations and curricula. In this way, validity supports the local and site-based assessment practices I discuss in chapter four, since "validity must be established for each particular use of a test" (Shephard 1993, 406).

In answering the question "What does it mean to say that actions based on test scores must be supported by evidence?" Lorrie Shephard (1993 406) uses the example of school readiness testing whose results are used to make some kindergartners wait a year before beginning their formal education. Shephard claims that in order to be valid, these decisions should be based upon evidence, showing that children profit from sitting out an extra year. In fact, no such evidence exists, with comparative studies even "show[ing] no academic benefit and some emotional harm" (407). This example about validity and readiness testing has strong implications for writing assessment, whether we refer to my earlier example of the placement testing at CUNY upon which the decision was made to deny entrance to certain students, or to the placement testing common at many colleges and universities that requires some students to take remedial courses before they can enroll in regular, credit-bearing writing

courses. In order to supply sufficient evidence about the validity of writing placement programs, we need to know how well those students denied entrance to a certain institution, or placed in a remedial class, ended up doing as a result of the decision that was based upon our placement procedures. Much more information is needed than is currently supplied by those who would consider our writing placement programs "valid."

In addition to requiring more evidence for claims of validity, there are additional differences between the way that college writing assessment has defined and talked about validity and the ways in which validity has been used by the educational assessment measurement community to make validity claims for its writing assessments. In either case, validity as it has been theorized is not the same as the practice used by either camp in writing assessment to justify the use of its assessments. I think it's possible and potentially very beneficial to view validity not as some pronouncement of approval but rather as an ongoing process of critical reflection (Moss 1998). In this way, as Moss and others (Bourdieu and Wacquant 1992; Cherryholmes 1988) advocate, validity is a way that "the inquiry lens is turned back on researchers and program developers themselves as stakeholders, encouraging critical reflection about their own theories and practices" (Moss 1998, 119).

CREATING A FIELD FOR WRITING ASSESSMENT

There is much to be gained for both the college writing assessment community and the educational measurement community if they would begin to use validity together as a way not only of regulating themselves and their assessments but also of developing assessments upon which decisions about writing can best be made. We have seen throughout this chapter that the isolation and conflict in writing assessment has been characterized in different ways, depending upon who tells the story. Probably the tension between the two camps in writing assessment can best be summarized as a conflict between values, between the need to produce consistent and replicable scores

in an efficient manner and the need to represent the complexity and variety inherent in written communication. A few years ago, I was part of a group helping to develop the writing test for the National Assessment of Educational Progress (NAEP). Our group was made up primarily of people who either worked in college writing assessment or were otherwise connected to the teaching of English. At one of the first meetings, we were being introduced to the new NAEP writing assessment by the psychometrician who was overseeing the project. He told us the parameters for the writing portion of NAEP, and how the budget was tighter than for the last writing assessment, and that even though he saw real value in projects like the 1992 portfolio portion of the previous writing assessment, this time, there just wouldn't be the resources for bells and whistles; he said that they were hoping for mostly single-sample twenty minute essays. Most of us who worked in English rather than measurement were a little stunned. However, James Marshall, who teaches in the School of Education at the University of Iowa put it best. "Your bells and whistles," he said "are our meat and potatoes." Much of the work we did as a group over an eighteen-month period could probably be characterized as trying to explain how what the educational measurement community considered to be fringe or extra accessories was for the English teaching community the heart of assessing student writing. In this situation, as in most of writing assessment conducted outside of the college writing community, the measurement people were clearly in charge, and most of the NAEP writing assessment went off as it had been initially planned by the personnel overseeing the project, regardless of much feedback to the contrary.

Including theoretical input about the complexity and context necessary to adequately represent written communication as a part of the validity process gives writing teachers and writing program administrators a real say about not only the ways in which student writing is assessed, but also the ways it is defined and valued. Of course, this does not mean that validity is an easy way for college writing assessment to take over writing assess-

ment as a field of study. While it allows the English teaching community a greater say in writing assessment, it also imparts other responsibilities. Validity inquiry requires what the educational measurement community calls "rival hypothesis testing," a process in which alternative explanations from both theoretical and empirical sources must be offered as well as alternative decisions based on the evidence. This consideration of rival explanations and actions is a central part of validity: "any validation effort still consists of stating hypotheses and challenging them by seeking evidence to the contrary" (Shephard 417). The process of considering rival explanations and actions is probably a sound method for any kind of serious thought. In writing assessment, however, it might be particularly crucial because it is a field in which, as we have seen, two competing communities are ready to advance different explanations for existing phenomena and different ways of gathering information to make important decisions about literacy education. Any validity inquiry in writing assessment, then, needs to include a serious consideration of rival theories, methods, explanations and actions, so that it includes a consideration of the values, ideas and explanations possible from both camps.

Lee Cronbach, a major figure in validity theory, characterizes validity and the act of validation as argument: "Validation speaks to a diverse and potentially critical audience; therefore, the argument must link concepts, evidence, social and personal consequences and values" (1988, 4). Two things make Cronbach's notion of validity as argument especially pertinent to writing assessment. One, his idea that validation documentation and research needs to speak to "a diverse and potentially critical audience" could not be more true considering our discussion of the two major camps in writing assessment. His point also highlights the necessity of building validity arguments that speak not only to those who share our disciplinary allegiances and theoretical and epistemological orientations, but to those who don't, as well. This imperative to use validity to cross disciplinary boundaries is crucial if we are going to work against the isolation

between college writing assessment and educational measurement and create a real field for the theory and practice of writing assessment.

The concept of the stakeholder is common in educational measurement, and it has been used by some in college writing assessment (White 1994), as we discussed earlier, to note the various kinds of people with an interest in writing assessment and the positions they hold. Conceiving of validity as a way to convince those who do not hold similar positions seems a significant way to account for difference in writing assessment. Using rival-hypothesis testing can make our arguments more palpable for a wide range of audiences. I see this concept in constructing validity arguments as way to work against a notion of writing assessment as dominated by distinct stakeholders with claims for varying degrees of attention to different theories and practices.

Trying to construct writing assessments that honor the legitimate claims of various stakeholders can result not only in the missed opportunity to create an assessment that can enhance teaching and learning, but it can also build assessments that are ultimately failures. The notion of honoring stakeholder's claims also ignores the politics of power. All stakeholders are not equal, and all claims will not and practically speaking cannot be equally honored. The need for technical specifications (Breland 1996; Camp 1996; Koretz 1993; Scharton 1996) or political control (Huot and Williamsom 1997) is often seen as more important than theoretical knowledge from the content area (Cherry and Witte 1998) or the needs and concerns of teachers (Callahan 1997, 1999) and students (Moss 1996; Spaulding and Cummings 1998). In writing assessment, the results of this unequal power struggle have been practices which score portfolios paper by paper to achieve interrater reliability (Nystrand, Cohen, and Dowling 1993) or portfolio systems that please neither the teachers (Callahan 1997, 1999), the students (Spaulding and Cummings 1998), school administrators, or politicians. Instead of attempting to honor disparate claims of unequal influence, we need to build writing assessment practices that have a firm

content-area theoretical basis and the potential to enhance teaching and learning. Emphasizing that validity addresses the decisions made on behalf of an assessment can only increase the importance of stakeholders like teachers, their immediate supervisors and students themselves, since it is these people who are most knowledgeable about the local educational process. Privileging the roles of teachers and students makes sure that assessment does not overshadow educational concerns.

In considering testing company personnel, I emphasize the role of content-area specialists like teachers and scholars from the supporting disciplines because their concerns are not usually considered in most large-scale or high stakes assessments. I also emphasize the role of content-area personnel in writing assessment because educational measurement specialists with no credentials in writing, rhetoric, linguistics or language education are not the best equipped people to integrate pedagogical implications in a writing assessment. Writing assessment as a contested intellectual site is an anomaly in educational measurement, since most other content-area fields do not have an active role in their assessments. Although I think it important to give additional power and responsibility to content-area professionals in writing, this responsibility also includes the necessity of constructing strong validity arguments. Any use of any writing assessment should be accompanied with a validity argument that addresses technical documentation important to those who work in educational measurement, honors political considerations important to administrative and governmental agencies, and most importantly considers the impact on the educational environment and the consequences for individual students and teachers. If validity arguments that consider all possible explanations and evidence are constructed, then those with various positions in a writing assessment can be represented. However, given that the commitment of validity theorists like Cronbach (1988), Messick (1989a, 1989b), Moss (1992), and Shephard (1993) clearly outlines the importance of assessment in creating environments conducive to teaching and learning, it follows

that if we are committed to assessments that promote teaching and learning, then we must listen primarily to the voice of educators and their students.

The second important aspect of Cronbach's characterization of validity as argument is that many of those in college writing assessment have a specific connection to the study of argument, since rhetorical study is an important resource for the field of composition. Not only does validity as argument pose more of an interest to those with a strong sense of rhetoric, it also give them a rhetorical heuristic for learning to construct validity arguments that contain a strong consideration of alternate views as well as an understanding of how to create arguments that are compelling to various audiences. Validity as argument provides the possibility for people who work in English departments and teach writing but are isolated from the literature and discipline of educational measurement to see validity as something familiar, understandable and valuable. White (1994a) has urged those interested in or responsible for writing assessment to become more knowledgeable about statistics and technical testing concepts like reliability, while at the same time promoting too simplistic an understanding of validity. I contend that college writing assessment and English teachers are better served by a current knowledge of validity theory. Validity in its rhetorical sense provides a way for college writing assessment to connect its assessment theories, scholarship, literature and practices to those in educational measurement. Of course, part of the rhetorical assignment college writing assessment developers undertake is to learn more about what the audience of those in educational measurement value if they are to be able to write validity arguments that convince educational measurement scholars. If we can promote the regular use of validity arguments that attempt to be compelling for all of those who work in writing assessment, then it might be possible to ease the current climate of isolation, since both camps in writing assessment would need to know about each other in order to make convincing arguments for validity.

In concluding this chapter, I hope to able to outline some new ways in which writing assessment can be understood as a field of study. It is clear that so far, writing assessment has been carried out by two different groups of scholars with different theoretical, epistemological and disciplinary orientations. Neither of the two camps has understood that the other is capable of enriching not only both points of view but writing assessment as a whole. In minimizing each other's contributions to writing assessment, each group has advanced its own impoverished version of writing assessment theory and practice. There are legitimate arguments from each side. College writing assessment can claim that the educational measurement community has advanced assessments that are not only ignorant of the ways in which people learn to read and write, but that these assessments have had deleterious effects on individual students and whole writing programs. Educational measurement can claim that college writing assessment not only appropriated measurement concepts, techniques and practices without acknowledging their origins but even ultimately misused them.

Validity, in its broadest and most current sense, can be a rallying point for both college writing assessment and educational measurement. Validity that looks not just at technical and statistical explanations but that focuses on the decisions and the consequences of those decisions made on behalf of an assessment cannot but help to appeal to those in college writing assessment. Validity as we have been discussing it and as the literature in educational measurement has been detailing for the last three decades has much to recommend it to the college writing assessment community. Stipulating that all claims for validity must consider theoretical and empirical evidence provides an opportunity for college writing assessment specialists to become full partners with their educational measurement counterparts. As I discuss in chapter six, reconceptualizing writing assessment as research rather than as a technical apparatus provides new leadership roles for teachers and administrators. Validity also imparts new responsibilities for college writing assessment,

since even if a department decides to use commercially prepared writing assessments, it is their obligation to provide a validity argument for each use of a test. Conversely, educational measurement scholars must begin to recognize the site-based, locally controlled assessments that are now being developed at many institutions. What I hope is that not only will those in educational measurement begin to recognize these assessments but that they will begin to help those in college writing assessment improve them and the validity arguments constructed for them.[2] Clearly, no matter which version of the field we subscribe to, there is much work to be done in writing assessment, and to accomplish this work, we need to draw on all the resources we have at our disposal. Creating a field of writing assessment that promotes communication, dialogue and debate can only increase our knowledge and understanding and improve the assessments we can create.