6

WRITING ASSESSMENT AS TECHNOLOGY AND RESEARCH

Several years ago, in an essay for a special issue in *Computers and Composition* on the electronic portfolio, I explored the ways in which technology had been applied to assessment, advocating that we use technology to link people together and to mediate responses to students rather than implement technology as a way to score student writing or respond to student writers. Writing that essay impressed upon me the ways in which technology had not only been applied to assessment, but how in many ways assessment had become a technology in and of itself. I later discovered that George Madhaus[1] (1993) had already been talking about assessment as technology and the problems this had brought with it:

> Changes over the last two centuries in the predominant ways of examining student achievement—from the oral mode, to the written essay, to the short-answer form, to the multiple choice format, to the machine-scorable answer sheet and finally to computer-adaptive—have all been geared toward increasing efficiency and making the assessment systems more manageable, standardized, easy to administer, objective, reliable, comparable and inexpensive. (82)

According to Madhaus, testing as we now know it is largely a creation of the twentieth century, in which social scientists eager to achieve scientific status for their work applied statistical explanations and technological apparatus to social and psychological phenomena like intelligence and aptitude. Along

with the statistical machinery of psychometrics, testing was also pushed toward a technological approach since there was an ever-increasing pressure to develop means to test the largest number of people in the shortest amount of time for the least possible money. While this pressure for efficiency was also motivated by the prevalent theoretical orientation toward education and reality (Williamson 1994), it was also motivated in practical terms by situations like the need to classify army recruits for WW I—for which the first truly large-scale assessment, the Army Alpha Test, was developed. The ability of the Army Alpha exams to classify nearly two million recruits bolstered the confidence of educational testers who during the twenties devised the SAT exam and other measures that led to the development of multiple choice testing and eventually machine scoreable answer sheets. By the end of the Second World War, the testing machine was incorporated with the establishment of ETS in 1947.[2]

For writing assessment, the technological focus is perhaps a little less clear, given the kind of assessment that actually involves the reading of student writing, that has been the focus of this volume. The most obvious manifestation of technology in writing assessment has been the euphemistically dubbed "indirect" tests—multiple choice exams of grammar, usage and mechanics, which are unfortunately alive and well. For example, the COMPASS test currently marketed by American College Testing (ACT) for college placement contains no writing at all, measuring only how well students edit a passage on computer. It has been given to over 750,000 students during the last few years. ACT's claims for the validity of the COMPASS are based on the same criteria used for other indirect tests of writing (see the discussion of Camp's explanation in chapter two for more details), namely that editing is part of the domain of skills necessary for writing and that there is a strong enough correlation between COMPASS scores and scores on essay exams in studies undertaken by ACT. Historically, writing assessment's technological focus has been fueled by its continuing emphasis on the technical problem of providing high enough rates of interrater reliability.

Since at least 1912 (Starch and Elliott) the use of essays to measure student writing ability was considered suspect because of the lack of agreement on scores from independent readers. The College Entrance Examination Board continued the use of essay-type tests into the 1940s. In 1937, the Board introduced the Scholastic Achievement Tests (SAT) as an experiment for students wishing to compete for scholarships. This SAT could be administered in one day, and was scheduled in April. This allowed colleges and universities to receive information on applicants much earlier than the more traditional essay exams, which were scheduled in June, after school commencement and which took an entire week to administer. By the early 1940s, the April examinations had grown in popularity. The Board had lifted the scholarship restrictions in 1939, allowing any college-bound student to take the exam. The number of students taking the April exams grew, while the numbers for the June examinations shrank. In 1942, the Board announced a series of new policies for examinations that were designed to aid the rapid enrollment and matriculation of students during America's wartime. Among these policies was the complete abolition of essay testing (Fuess 1967). In response to a strong backlash against the scrapping of essay exams, the Board instituted a one-hour English Composition Test in 1943 as one of its achievement tests (Fuess 1967). Eventually, the use of indirect measures solved the interrater reliability problem, since these exams did not involve the reading and scoring of student writing. In the 1950s, interlinear exercises were developed in which students were to correct a text that contained specified errors (the format of the COMPASS). By the 1960s, a team of ETS researchers (Godshalk, Swineford and Coffman 1966) were able to devise methods that ensured enough agreement among researchers to make the scoring of essays statistically viable. Overall, methods for writing assessment have evolved from reading essays to make a specific decision, to multiple choice tests, to holistic, analytic and primary trait scoring, to computer scoring of student writing. As various technologies were developed throughout the

twentieth century, they were applied to writing assessment. Now, it is possible for students wishing entry to graduate school in business to write exams on computer from all over the world, and for readers to receive training, score essays and relay their decisions by accessing a secure web site.

Clearly, there are some advantages technology can provide for the assessing of student writing. For example, writing placement has always presented real problems, since students have to write an essay which needs to be scored, so they can be placed into a specific class. Often the first time students are available to write is when they visit campus during orientation; this dictates that a tremendous number of student essays need to be read within a short time, so students can be registered for the appropriate class. Technology now permits students to compose on computer or even online, with readers accessing student writing from a secure web site, expediting the entire process—which greatly aids students and schools who need such decisions as soon as possible.

WRITING ASSESSMENT AS TECHNOLOGY

Understanding writing assessment as technology is important because it gives us a lens through which to consider the ways in which assessment procedures have evolved. Also, it's important to keep in mind that technologies can be imbued with various political and ideological orientations. Perhaps the most famous example of technology and ideology comes from Langdon Winner (1986) who points out that the underpasses on the Long Island Expressway were deliberately designed so that busses would not fit through them. The idea behind such a design was that it would keep people off the island who could not afford their own vehicles and relied on public transportation. As I mention in the next chapter, the Standford Binet IQ test was actually renormed in the 1930s after initial versions showed that girls outperformed boys. Technological projects like intelligence testing or engineering have the veneer of being objective, scientific and socially disinterested, but as studies in the rhetoric of science and these two examples demon-

strate, all human activity is situated in specific ideas about society and social order, and all professional practices have theoretical, epistemological and material consequences. Technology, like any other human activity, can help to promote certain social, political and ideological values.

In his book, *Critical Theory of Technology*, Andrew Feenberg identifies two main attitudes toward technology: instrumental and substantive. The less common substantive attitude toward technology refers to people who believe that "technology constitutes a new type of social system that restructures the entire social world as an object of control" (1991, 7). Instrumental, the more common attitude toward technology, the one which sums up the way in which assessment and technology have been linked together, stipulates that "Technologies are 'tools' standing ready to serve the purposes of its users. Technology is deemed neutral without valuative content of its own" (1991, 5). In this way, technology can be used in many different ways merely as a tool to accomplish a particular task. Unfortunately, this instrumental view of technology as an ideology-free problem-solving tool can lead to an approach Seymour Papert (1987) called "technocratic" in which computers and other technologies are viewed outside of any specific context. The ability to accomplish a task or solve a problem is merely dependent upon our ability to devise an appropriate technological solution based upon the available technology. In other words, we do something because we can. Gail Hawisher (1989) adapted Papert's notion of the technocratic in coining the term *technocentrism* to describe the enthusiastic and uncritical employment of computer technologies to teach writing. Hawisher argues that in the rush to embrace computer technology for the teaching of writing, we abandoned what we knew about literacy, composing, and reading, focusing instead upon the kinds of instructional environments that computers made available, regardless of the ability of these environments to foster the qualities we know enhance the teaching of writing. In a technocentric approach, the ability of the tools themselves drive our practices.

For example, in his germinal chapter on holistic scoring in 1977, Charles Cooper introduces holistic scoring as a way of assessing student writing, no small feat considering that before the use of holistic scoring, using essays to assess student writing was considered unreliable, and multiple choice tests of grammar usage and mechanics were the only statistically acceptable form of writing assessment. Cooper goes on to expound the virtues of holistic scoring, noting that now we could now rank the writing abilities of every student in a school. Cooper provides no reason for why we would want to rank students in a particular school beyond the fact that it is now technically possible. Cooper's reasoning could be said to be technocentric, since it appears to be based on the fact that holistic scoring supplies school officials with the ability to rank students rather than on any other consideration of the ways writing is taught or learned or how this ranking could improve teaching or learning. What's important to remember is that while Cooper's idea to rank all students in a school based upon their scores from holistic scoring comes from a technocentric view of assessment, the results of such a decision could have widespread implications for teaching and learning of writing, depending upon who has access to the rankings and how they might be used to make educational decisions.

Cooper's inclination to use the newly developed technology of holistic scoring to evaluate and rank all the students in a school is representative of the ways in which writing assessment has been developed over the years. Because early studies of writing assessment showed that independent raters had trouble agreeing on what scores to give the same essays, writing assessment focused on how to achieve reliability, to the point of developing tests that were reliable even though they contained no student writing at all. Frank Bowles, president of the College Board called the writing sample portion of the entrance examination "an intellectually indefensible monstrosity" (Valentine 1980, 116). During the backlash the Board experienced after dismantling the essay exams, John Stalnaker, the Board's Associate Secretary writes in 1943:

> The type of test so highly valued by teachers of English, which requires the candidate to write a theme or essay, is not a worthwhile testing device. Whether or not the writing of essays as a means of teaching writing deserves the place it has in the secondary school curriculum may be equally questioned. Eventually, it is hoped, sufficient evidence may be accumulated to outlaw forever the "write-a-theme-on" . . . type of examination. (qtd in Fuess 1967, 158)

Stalnaker's blistering criticism of essay testing is less shocking than his complete indictment of teaching writing by having students write. For me, what's even more disturbing is that his words seem to have had great predictive value, given Arthur Applebee's findings in the late 1970s about the absence of writing in the secondary curriculum (1981).

In trying to account for why it took so long for the CEEB to move toward the indirect testing of writing through the use of multiple choice tests on grammar usage and mechanics, Orville Palmer of ETS writes,

> The Board regretted the authority of a large and conservative segment of the English teaching profession which sincerely believed that the writing of essays and other free response exercises constituted the only direct means of obtaining evidence as to a student's ability to write and understand his own language. (1960, 11)

Palmer's history of the "Sixty Years of English Testing" goes on to elaborate how "more complex testing techniques" were eventually developed.

The use of a multiple-choice test to assess student ability in writing certainly fits even a simple definition of technology as "something put together for a purpose, to satisfy a pressing and immediate need, or to solve a problem" (Madhaus 1993, 12-13). According to Madhaus, the notion of technology as a machine has evolved into a newer conception: "However, much of present technology is specialized arcane knowledge, hidden algorithms, and technical art; it is a complex of standardized means for attaining a predetermined end in social, economic, administrative and educational institutions" (12). In this way, it is possible to

see reliability as a technical problem and the use of multiple choice exams of writing as a technology to solve this problem, involving the "arcane knowledge" of test item analysis, concurrent validity and psychometrics. In other words, a multiple choice test becomes a viable way of assessing writing because it is technologically possible, satisfying the technical need for reliability, though it may not contain any writing. Writing assessment has been predominantly constructed as a technical problem requiring a technological solution. Donald Schön's discussion of positivism and Technical Rationality, including his comment that "professional activity consists in instrumental problem solving made rigorous by the application of scientific theory and technique" (1983, 21), provides a good description and explanation of the ways in which writing assessment developed as a technology. For Schön, the importance of this problem-solving orientation is especially crucial in how knowledge gets made:

> In real-world practice, problems do not present themselves to the practitioner as givens. They must be constructed from the materials of problematic situations which are puzzling, troubling and uncertain . . . But with this emphasis on problem solving, we ignore problem setting . . . Problem setting is a process in which interactively, we *name* the things to which we will attend and *frame* the context in which we will attend to them. (1983, 40 [italics in original])

In other words, by setting up the technical problem of reliability as the main agenda for developing writing assessment, early writing assessment researchers ignored the way in which the problem was set and instead focused on how to create procedures for reading and scoring student writing in which teachers could agree. More recent writing assessment procedures (Durst Roemer and Schultz 1994; Haswell and Wyche-Smith 1994; Smith 1993; and others), which we have discussed throughout the rest of this volume, circumvent the focus on interrater reliability in various ways. Durst, Roemer and Shultz, for example, have teachers read in teams in which they discuss their decisions. Haswell and Wyche-Smith only use one reader for the initial reading.

Smith paired readers with similar teaching experiences and found that they agreed at a higher rate than those he trained with conventional holistic scoring methods. Had early writing assessment specialists followed Smith, they might not have discovered any reliability problem at all, since Smith's results indicate that when teachers make contextual decisions about which they are expert, they tend to agree at a higher level than explicit training for agreement can provide. One main difference between conventional writing assessment procedures and those I cite here is that holistic, analytic or primary-trait all render scores for writing; whereas, all of the newer forms render direct decisions about student writers based upon their writing. Technological approaches produce uniform, standardized and abstract results like scores, whereas newer approaches produce direct, concrete, contextual and applicable decisions.

The technological focus of writing assessment can be seen in the recent and continuing creation of computer programs that can "simulate" human readings by providing the same score a trained reader might give the same essay. The emphasis on reliability that first led to the development of indirect tests of writing and then holistic, primary-trait and analytic scoring is now leading the development of computer programs to generate reliable scores. In order for a writing assessment scoring session to be deemed acceptable, it must display an acceptable level of interrater reliability, represented as a numerical coefficient computed according to specific standard statistical formulae (Cherry and Meyer 1993). Test administrators and those who read essays for holistic, analytic or primary-trait scoring must follow specific procedures for training, reading, and scoring student papers[3] or portfolios. Within these scoring sessions that are geared to providing consistency in scoring, readers are often asked to suspend their own reading of student writing in order to read according to the guidelines specified in the rubric, so that raters can agree (Broad 1994; Bunch and Schneider 1993). Training readers is sometimes referred to as a calibration process, as if readers, like some machine, are calibrated to agree

(Bunch and Schneider 1993; Hake 1986). As White (1994) and others have detailed, holistic scoring[4] requires adherence to a fairly tight protocol of procedures.

The main result of this tightly scripted procedure for reading student writing is the production of reliable scores for writing assessment. Other results include the creation of a reading system that favors one particular interpretation of reading and student writing ability. There is no room for the diversity of opinion and interpretation that mark most reading (Broad 1994; Elbow and Yancey 1994; Williamson 1993). In a description of a training session, Robert Broad (1994) illustrates the ways in which interpretation and meaning are predetermined through the use of anchor papers and training rubrics, so that readers are compelled to read student writing from one specific point of view. More recently, Pamela Moss and Aaron Schutz (2001) illustrate the same phenomenon in describing the creation of standards for teacher certification. This way of reading is antithetical to the kinds of meaning making and interpretation that most often accompanies the way in which people come to value certain kinds of texts (Smith 1988). As Elbow and Yancey (1994) note, reading in a scoring session designed for agreement alters the focus present in much reading in English, since different, innovative and novel interpretations are valued in reading literature, whereas in reading student writing, such diversity is all but abolished. In fact, Edward Wolfe, in a series of research studies, has found that raters who agree most often with others in scoring sessions actually read in a more limited and focused manner that emulates the principles fostered in training and in the scoring rubric (Wolfe 1997; Wolfe, Kao, and Ranney 1998; Wolfe and Ranney 1996).

It is safe to say that holistic and other scoring methods that rely on rubrics and training and are designed specifically to foster agreement among raters produce an environment for reading that is unlike any in which most of us ever read. What's more crucial in understanding the machine-like orientation of holistic, analytic and primary-trait scoring is that little attention is

directed toward anything else. For example, Englehard, Gordon, and Gabrielson (1992) conducted a study involving the holistic scoring of over one hundred thousand pieces of student writing and used this to make important points about gender and writing ability, even though the scoring rubric is predominantly focused on usage, grammar and mechanics. These kinds of scoring procedures are best understood as a technology designed to solve the problem of inconsistency in scoring. The result of this technological approach to writing assessment is that people choose to conduct holistic scoring sessions to produce numerical scores for students regardless of the decisions they wish to make. What's probably even more problematic is that attention during and after a holistic scoring session is focused on the technical aspects of scoring, like the construction of the scoring guideline or rubric, the selection of anchor papers, and, most importantly, the generation of reliable scores. Writing assessment practitioners are more like technicians than anything else as they attend to the machinery of the scoring session, since these are the important aspects that will make the assessments acceptable and valuable. Scoring sessions are standardized routines whose very acceptability depends upon the strict adherence to certain procedures that were designed to ensure the reliable production of scores for student essays.

The overwhelming technological focus of writing assessment has created a climate in which technical expertise is continually emphasized, and attempts to create assessments outside of a narrow technological focus are severely criticized: "Authorizing English departments to isolate themselves intellectually in order to engage in technically amateurish evaluation of their programs" (Scharton 1996, 61). In fact, English teachers who refuse to support traditional, technological approaches to writing assessment like the computer-generated assessment of student writing have been grossly caricatured: "A political stance that denies the importance of writing mechanics and resists all forms of technology and science is not good for writing instruction" (Breland 1996, 256). While Maurice Scharton's (1996)

claim of isolation for English departments mirrors Pamela Moss's claim, which I discussed in chapter two, that college writing assessment remains isolated from the larger educational community, my point that the isolation Moss refers to goes both ways seems equally true for Scharton's (1996) claim. For example, he fails to cite several of the writing assessment programs developed in English departments—programs that I discuss throughout this volume. Not only do assessments like those developed by William L. Smith (1993), Richard Haswell and Susan Wyche-Smith (1994), Russel Durst, Marjorie Roemer and Lucille Schwartz (1994) among others defy the label "amateurish," they have, as I illustrate in chapter four, been able to break new ground, providing interesting and innovative approaches outside of current traditional writing assessment. Hunter Breland's (1996) description of English teacher's attitudes toward assessment, correctness and science has no basis in reality. Perhaps one positive way to understand these and other attacks by members of the assessment community is to see them as a desperate attempt to combat the eroding influence of their technological focus, clearly a signal that important contributions are currently being made by the college writing assessment community and others working outside a narrowly defined psychometric and technological focus.

WRITING ASSESSMENT AS RESEARCH

The biggest problem created by the way writing assessment has been developed and constructed as a technology primarily to solve the technical problem of interrater reliability is that it has obscured the essential purpose of assessment as research and inquiry, as a way of asking and answering questions about students' writing and the programs designed to teach students to write. The primary consideration in assessing student writing should be what we want to know about our students. For example, are they ready for a specific level of instruction, or have they completed a course of study that allows them to move on to new courses and challenges? When doing research, the primary

considerations are the research questions. Once we decide what it is we want to know, then we can fashion methods to help us find out. With the current conception of writing assessment as a technological process, the methods themselves have become the most important consideration. Procedures like the construction of a scoring guide or rubric and the training of raters on that rubric have become more than just research methods. Rather than have the research questions drive the search for information, the methods themselves have become the focus of those who conduct writing assessments. The methods of writing assessment receive so much of our attention that they have, in effect, become reified as writing assessment itself.

Instead of devoting assessment to asking and answering questions about student writing or its teaching, those conducting assessment spend their time worrying about and perfecting the technical aspects of scoring student writing. Although we have, as Yancey (1999) notes, seen the evolution of writing assessment from multiple-choice exams to single essays to portfolios, most student writing is still assessed using holistic or other scoring methods that require rubrics, rater training and the like. In other words, we have changed the sample of student performance from answers on multiple-choice tests of usage and mechanics to multiple writing samples, but we are still using the same research methods supported by the same theoretical and epistemological orientation (see chapter four) to render decisions about students. This ongoing reliance on these specific methods for research continue to foster the technological values that led to their development in the first place. While the multiple texts in a portfolio provide an opportunity for recognizing various kinds of texts and student ability in writing those texts (Belanoff 1994; Berlin 1994), the technology of holistic scoring strives to render one true reading (Broad 1994, 1997, 2000; Elbow and Yancey 1994; Williamson 1994). If we were to change the focus of writing assessment from the use of specific methods to a process for inquiry, we would, in effect, be changing not only the ways in which writing assessment is conducted

but the culture surrounding assessment, the role of assessors and the products of our assessments, providing the possibility for real change in the ways we think about writing assessment and the positive role assessment can play in the teaching of writing and the administration of writing programs.

As I note in chapters three, four and five, writing assessment is not something most teachers see as related to or beneficial to their goals in teaching students how to write. In fact, many teachers see assessment as a negative force because so many current assessment practices do not even attempt to address teaching and learning, yet they nonetheless narrow or guide instruction, since many high stakes decisions about students, teachers and programs are linked to student performance on assessment measures. Expanding the methodological options for writing assessment and for the roles of teachers and program administrators not only furnishes the opportunity for us to collect and analyze a wealth of new information about student writing, it also provides additional motivation for teachers to become involved in writing assessment. With writing teachers in charge of assessment, there is the possibility that the culture surrounding assessment can be revised. As I note in the next chapter, there appears to be no cumulative body of knowledge among writing teachers and administrators concerning writing assessment. What writing assessment culture does exist often revolves around a sense of crisis, in which assessment is cobbled together at the last minute in response to an outside call that somehow puts a program at risk. At best, writing assessment is seen as something that we better do before it's done to us. Of course, this sense of assessment is understandable given the current nature of most assessment practices and initiatives—in which writing teachers and administrators are expected to use particular methods that require some technical and statistical skills most English teaching professionals do not possess. Envisioning writing assessment as research, however, alters the relationship between teaching, learning and assessment, since the teachers themselves are involved in articulating questions

about their students, programs, and teaching, and are designing methods to answer the questions they have actually posed. Questions about how well students are doing in specific classes and on specific assignments also become venues for teachers to talk about what they value in their teaching, their expectations for their students and their overall sense of how successful they and their colleagues have been in the teaching of writing.

If we see our task in writing assessment as research, it not only changes the focus of the activity, it also changes the role of the assessors. Instead of administering a pat set of procedures to produce student scores and interrater reliability, it is necessary to decide what a department or program wants to know about their students' writing, their teaching and the overall effectiveness of their writing program. Acting as researchers changes the role teachers and administrators might play because instead of just being technicians who administer the technological apparatus of holistic or other methods of scoring, writing teachers and program administrators become autonomous agents who articulate research questions and derive the methods to answer those questions. A few years ago, there was a popular movement in composition that advocated that teachers assume the role of researcher in their own classrooms and department. This teacher/researcher role can have new meaning for a brand of assessment that brings teachers together to articulate questions about how well students are writing and how well we are teaching them to write. This change in role accompanies a change in power, as well. Teachers and their assessments can no longer be judged on just how technically sophisticated they are (Camp 1996; Scharton 1996). I'm not saying that the technical aspects of assessment research are unimportant. However, I do not think they should be the primary concern. It is possible for those in English to receive technical assistance from colleagues in education, psychology, or measurement, creating the kinds of coalition and connections for a field of writing assessment that I advocate in chapter two. Instead of just focussing on its technical aspects, writing assessments will need to be evaluated on how

well they articulate specific questions, provide data to answer those questions and ultimately analyze that data to effect important changes in teaching and program administration.

In addition to giving teachers new agency for assessment, seeing assessment as research can also further alter the often inherently unequal power relations in assessment. While few English departments can boast of any expertise in writing assessment, many English and writing departments do include people with experience and expertise in research. Knowledge about and experience in asking questions and deriving methods to answer those questions is expertise that English teaching professionals and others can use to conduct writing assessments on their own campuses. Not only does encouraging literacy researchers to become involved with assessment enhance the position and role of English teachers conducting their own assessment, it also creates the real possibility that we can become aware of new ways to gather and analyze data to make important decisions about students and the teaching of writing. Currently, most writing assessment is conducted using holistic scoring procedures developed by ETS in the 1950s and 1960s. While the rest of composition studies has seen an explosion in the exploration of qualitative research methods that reflect new concerns about the social, situated nature of literacy and the political and ideological issues of representation and power, writing assessment methods lag behind. Except for the locally-developed methods I discuss in chapter four, current writing assessment research methods focus on how to produce reliable scores among readers.

Employing qualitative methods that appear to be suited for gathering and analyzing information about literacy and its teaching should also alter the products that assessments can produce. New methods of assessment that employ qualitative methodologies can provide thick descriptions of the kinds of writing instruction and performances that occur in our classrooms and programs. It might be possible, for example, to categorize various kinds of writing instruction and student performances, giving detailed examples of student and teacher performance. Instead

of just being able to say certain students can satisfy *x* number of outcomes or standards or other sorts of criteria that are easily assembled, quantified and aggregated, it might become possible for assessment to provide the site for rich, descriptive examples of student writing and development. In this way, we can draw upon the theory and practice in educational research that advocates a multimodal approach. Before this can happen, of course, it is necessary for us to begin to ask different kinds of questions in our assessments—questions that require us to collect different kinds of data and perform different sorts of analyses. Asking new questions, employing new methods, and using assessment as research are only a few of the ways in which assessment as research can transform what we now know as writing assessment.

It should be noted that current traditional writing assessments that are developed and practiced by reputable testing companies like ETS do involve much research. This research, however, is about the tests themselves—e.g., studying whether or not certain students respond to particular prompts in certain ways, or how student scores match up to grades or other indicators of achievement or aptitude. While it is commendable that companies research their tests, seeing writing assessment as research is fundamentally different; I am not talking about the research done on a specific test, but rather am advocating that the assessment itself be seen as research. So, while some conventional writing assessments can tout their research programs, they are still only providing a minimum amount of data on each specific student—usually the sum of two scores he or she receives. As current theories of validity advocate, writing assessment as research opens up the possibility that we collect various kinds of information about students before we can make important educational decisions. Test development research on technical matters is quite different from arguing that we must collect and analyze richer information to make important decisions about students.

Seeing writing assessment as research also gives us a powerful lens to view its development and history. For example, Yancey characterizes the history of writing assessment as series of waves:

> One way to historicize those changes is to think of them as occurring in overlapping waves, with one wave feeding into another but without completely displacing waves that came before. These trends marked by the waves are just that: trends that constitute a general forward movement, at least chronologically, but a movement that is composed of both kinds of waves, those that move forward and those that don't. (1999, 483)

According to Yancey, writing assessment history has not moved forward in any orderly fashion, so she offers the metaphor of successive waves. The wave metaphor allows her to describe how some things like multiple samples of student writing, part of a new wave, have changed, whereas holistic scoring, part of an older wave, has not changed. In this characterization of writing assessment, we don't always know what a new wave will bring or what will be left behind.

However, if we think about writing assessment as research, it may be possible to predict what will change and what will remain the same. For example, if we think of writing assessment as research, then we can separate writing assessment into the sample of what students produce and the way in which this sample is analyzed to make a decision. Yancey characterizes writing assessment history waves according to the following scheme:

> During the first wave (1950–1970), writing assessment took the form of objective tests; during the second (1970–1986), it took the form of the holistically scored essay; and during the current wave, the third (1986–present), it has taken the form of portfolio assessment and of programmatic assessment. (1999, 484)

In each of Yancey's waves, what changes is the sample of what students produce. For example, in the first wave, students provide information about their knowledge of usage, grammar and mechanics. This information is collected in a multiple-choice format, which is by definition reliable, and is used to make a decision about student's ability to write. In the second wave, a single sample of student writing is produced, and this sample is read by readers trained to agree on a specific scoring guideline,

to produce reliable scores (the same aim of the multiple-choice tests). In the third wave, the sample changes again, since students submit multiple samples of their writing, but the analysis of this writing can remain the same, since the two portfolio programs at Michigan and Miami that Yancey refers to use holistic scoring to arrive at decisions about the portfolios. What's apparent from her discussion is that what changes in each of the waves Yancey describes is the sample of student work, while the unit of analysis can remain the same. Yancey's focus on just the sample of student work and her lack of interest in the way the work is analyzed is even more apparent in her statement about the scoring system developed by William L. Smith, which she calls a "second wave holistic model" (1999, 496). Yet Smith (1993) actually developed and compared his expert reading system in opposition to the holistic method he was currently working with. Since Smith used a single sample of student writing, Yancey calls it holistic, even though Smith's method of analyzing the sample was radically different from holistic scoring because it contained no rubrics, rater training or other procedures associated with holistic scoring. If viewed solely in terms of the sample of student writing, the importance of Smith's groundbreaking work (which I refer to throughout this volume) is lost. Clearly, seeing writing assessment as research can be a powerful and important lens through which to view its development.

It is also important in a discussion of writing assessment as research to note that it is not enough to merely raise questions about the amount and quality of information we collect about students and how well we analyze that information to make decisions about that student and others in our programs. We must also ask questions about the methods we are using to conduct this research on assessment. That is, we must not only turn our research gaze outward toward our students and programs but inward toward the methods we are using to research and evaluate our students and programs. This kind of research is often referred to as validation. As I argue in chapter two, validity is often presented in a much simplified form in the college

assessment literature. Instead of just viewing validity as whether or not an assessment measures what it purports to measure (White 1994; Yancey 1999), it's important and necessary for us to consider the work of major validity theorists like Samuel Messick (1989a, 1989b) and Lee Cronbach (1988, 1989). Although I mention the importance of validity in chapter two and cite definitions offered by Messick (1989a) and Cronbach (1988) in chapter four, I would like to conclude this chapter by using the current conceptions of validity in an extended example to illustrate the important ways in which research can improve and transform writing assessment practice. In addition, it is crucial not only that we extend our understanding of assessment as research to include the data we collect and analyze to make decisions about students, but also that we take responsibility to research our own assessments.

THE VALIDATION PROCESS AS RESEARCH

One of the most interesting examples of writing assessment as research comes from the work of William L. Smith (1993), which I discuss in some detail in chapter four and in other places throughout this volume. Although Smith created the first placement system that did not rely on holistic scoring or other methods derived from measurement specialists and psychometric theories, he didn't start out to do this. In a discussion of writing assessment as research, what's important to note is that Smith's placement system was a result of several research studies he undertook to uncover what he perceived to be a forty percent error rate in placement. Smith's (1993) series of research studies permitted him to ascertain that his program did not in fact contain such a large rate of error. In addition, he was able to discover many other things about the way in which students were placed, how student essays were being read and how well students performed in the classes into which they were being placed. In a very real sense, what Smith did was to conduct validation research on his program with the ultimate end of revising completely the way in which his program placed students.

Smith's research illustrates that the process of validation research demands that we either supply information and analysis to support the decisions we're making about students or create new procedures that are supported by the data we collect and analyze.

I'm going to outline the kinds of validation research we conducted at the University of Louisville to justify the use of state-mandated portfolios to place students into first-year writing courses, but before I do that, I'd like to summarize our discussion of validity from chapter two and outline the specific course of validation I chose. Although Samuel Messick (1989, 1989a) and Lee Cronbach (1988, 1989) are the two most influential validity theorists of the last few decades, for our purposes here I would like to focus on Cronbach's ideas. For Cronbach, "Validation speaks to a diverse and potentially critical audience: therefore, the argument must link concepts, evidence, social and personal consequences and values" (1988, 4). Cronbach's notion of validity as argument seems to me particularly relevant for those who teach writing and are interested in its assessment, since it lends a rhetorical framework for establishing the validity of making decisions based upon specific writing assessments. While writing teachers and program administrators tend not to be knowledgeable about the technical aspects of measurement and validation, they are comfortable with and knowledgeable about the ways in which arguments can be crafted. Seeing validation as argument also illustrates some important features of validity.

For one, validity will always be partial, or as Messick notes, validity is "an evaluative judgment of the *degree* to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based upon test scores or other modes of assessment" (1989b, 13 [italics mine]). An argument is always partial since it is possible that some will be persuaded and others not. According to Lorrie Shephard, "In the context of test evaluation, Cronbach reminds us that construct validation cannot produce definitive conclusions and cannot ever be finished (1993, 430). The partiality of argument

and validity is based upon the truism that "What is persuasive depends upon the beliefs in a community" (Cronbach, 1989, 152). Richard Haswell and Susan McLeod (1997), in their chapter on writing-across-the-curriculum assessment, illustrate this by discussing the ways in which different academic and administrative audiences respond best to various types of reports and documents on program assessment. Argument and validation, then, necessitate that we consider audience. This notion of audience in argument is often seen as addressing the ideas of others, what others' arguments might be, and how we might persuade them anyway. In validity theory, a more formalized notion of this idea of audience consideration is called "rival hypothesis testing" which "requires exposing interpretations to counter explanations and designing studies in such a way that competing interpretations can be evaluated fairly" (Shephard 1993, 420). As well, we ask other audiences or constituencies for possible rival hypotheses in order to generate validation procedures that are as persuasive as possible. One last element of validation as argument is specificity. Just as arguments need to be specific in order to be persuasive, so too validation involves looking at a particular use of a specific measure:

> To call Test A valid or Test B invalid is illogical. Particular interpretations are what we validate. To say "the evidence to date is consistent with the interpretation" makes far better sense than saying, "the test has construct validity." Validation is a lengthy, even endless process. (Cronbach 1989, 151)

I am fairly certain that after reviewing some basic concepts about validity, our attempts to validate the use of state-mandated writing portfolios for placing first-year college students into courses will seem far from ideal. However, the process of validation is important in and of itself. Pamela Moss (1998) notes validation should be seen as a reflective practice through which assessment researchers can scrutinize their own efforts. It is in this spirit then, that I summarize the procedures that my colleagues and I (Hester, Huot, Neal and O'Neill 2000; Lowe

and Huot 1997) undertook over the course of eight years to validate the use of state-mandated portfolios for placing students into first-year writing courses at the University of Louisville.

We began the portfolio placement project as a pilot, at the request of the admissions director and a school of education faculty member who had contacted five schools in the Louisville area to see whether or not the students would be interested in using the senior portfolios they were required to submit as part of an overall assessment of the state's schools. The first year we read fifty portfolios. In subsequent years, we opened up the pilot to more and more students, and for the last five years, any student in the State of Kentucky has had the option of submitting her senior portfolio for placement purposes. In the last four years, we have averaged around 250 students, more than ten percent of the incoming class. Having at least ten percent of the incoming class participate in the project was important to us because without a substantial number of students participating, we felt very limited in terms of what we could claim for the process itself. Since our goal was to establish portfolio placement as the standard practice for placement, we would need to know how well it worked with a quantity of students. On the other hand, we were/are aware that our pilot was a self-selected sample, and that students who participated in the pilot would not necessarily reflect the majority of students who enroll in our classes.

Our first question was whether or not we could use high school portfolios written for a specific assessment program for another purpose like placement. We resisted the Department of Education's offer to have our readers "trained" on the state's rubric which has shown that these portfolios can be read reliably. Instead, we modeled our reading procedures on those developed by William L. Smith (1993), in which readers are chosen for their expertise in certain courses. Over the seven years we conducted the pilot study, we have continued to revise Smith's procedures. Unlike Smith, we allowed just one reader to make a placement decision. Our other revisions mainly consisted of streamlining the process. Eventually, we came to have all portfolios read first

by English 101 readers, since most students receive that placement. Then, we asked readers who are expert in the courses above and below 101 to read. In this way, our procedures emulate those developed by Richard Haswell and Susan Wyche-Smith (1994), who demonstrated that most students can be placed by a single reading into the most heavily enrolled course.

The first question we needed to answer if we were going to use portfolios for placement was whether or not we could get students into appropriate classes. We answered this question in two ways. Initially, we kept records of how well students did in the courses into which they were placed (Lowe and Huot 1997). For the last four years, we have also asked instructors to tell us if their students were accurately placed into their courses. We surveyed teachers from all classes about all students, so instructors had no idea which students were placed by portfolio. Additionally, we asked students about their placement and the use of their portfolio for placement purposes. (Hester, Huot, Neal, and O'Neill 2000).

Because this was a pilot project, for the first five years students who submitted portfolios for placement also went through the existing process for placement. This allowed us to compare the pilot use of portfolios with existing procedures, which consisted of using students' ACT score as an indicator of whether or not they need to write an impromptu essay. If students scored lower than eighteen[5] on the verbal section of the ACT, they were required to write an impromptu essay. We compared student placement based on the two different procedures as well as comparing how well students from different placements did in their first-year writing courses. Although we have found that portfolios tend to place students higher than the use of an impromptu essay does, we have also found that students with a portfolio placement achieve as well as those placed with existing placement procedures. We have further found that both instructors and students are happy with portfolio placement.

Because we do not require that each portfolio be read by two different readers, we cannot report interrater reliability data for

all portfolios. We did, however, compile interrater reliability data for all portfolios that were read twice, and found that the level of agreement was equal to or higher than (one year it was 100%) what is normally accepted. Because the reliability of an instrument is an important component of validity (Moss 1994a), we began checking for the reliability of placement decisions from one year to the next. Readers are either given a set of papers which had been read and placed the year before or, as we have begun to do in the last couple of years, a small percentage of portfolios are read twice even though the student can be placed on one reading. We have found over the last four years that the degree of consistency is more than acceptable.

Although students from the first five years of the portfolio placement project were required to complete both placement procedures, they were also allowed to chose which placement they wanted. We have documented the choices students made and the results of these choices in terms of the grades they received and the level of satisfaction they and their instructors expressed about their placement. So far, we have found that students always chose the higher placement and that for the most part they were successful in the courses they chose. Although our original intent was just to find out how well portfolios would work for placement, we hope to use this information in designing future placements in which students can have the opportunity to make choices about where courses most fit their needs. Currently, Directed Self Placement (DSP) is becoming a popular option for many schools who see the value of allowing students to make to make their own decisions about first-year college writing placement (Royer and Gilles 1998). Unlike Royer and Gilles's system, however, students at Louisville who made choices relative to placement did so after having received recommendations based on their own writing. This seems to answer criticisms about DSP that suggest students do not have enough information upon which to base their decisions and otherwise rely on gender or other stereotypes, and that DSP lacks enough empirical evidence for its claims (Schendel and O'Neill 1999).

Our validation data even includes information about the cost of reading portfolios. Messick (1989) makes the case that validity should include whether or not to undertake an assessment. Writing assessment that actually includes the reading of student writing is much more expensive than a multiple-choice exam, because of the additional costs of paying readers. In fact, the move away from essay exams in the 1930s by the College Board was partly predicated on the fact that the newer exams were less costly. As portfolio writing assessment became more and more popular in the 1990s, Edward White (1995) and I (1994) both questioned whether or not portfolios for placement were worth the additional cost. However, during the Louisville portfolio placement project, we have been able to increase the hourly rate we pay readers and still manage to hold the lid on costs, mostly because many of the changes we implemented streamline the reading process. Over the eight years we have read portfolios, we have never averaged more than five dollars per portfolio, and in the last couple of years we have reduced the cost to a little over three dollars.

Certainly our validation procedures for the portfolio placement pilot have not been as extensive nor have they rendered as dramatic results as those undertaken by William L. Smith (1993). However, the process allowed us to make a convincing argument that has resulted in the University of Louisville's decision to accept portfolios as a regular way for students to achieve placement in first-year writing classes. The process, as I hope I have demonstrated, not only allowed us to make a case for portfolio placement, but it also allowed us to learn more about the best way to conduct portfolio placement on our campus. And, like Smith, we also learned some surprising things along the way in terms of students choosing their own placement, a concept none of us had ever heard about before we undertook the pilot program and its validation. This last point underscores what is perhaps a crucial distinction between assessment as technology and assessment as research. Following prescribed methods designed to produce reliable scores for student writing is probably never

going to provide the opportunities for making knowledge that engaging in inquiry-driven research can often give us.

CONCLUSION

In truth, assessment has always been just another kind of research designed to provide us with information about student performance or the performance of the programs we design to help students learn. As Moss (1994a) and Williamson (1994) have pointed out, the need to attain reliable assessment has pushed for more and more standardization, until assessment, as Madhaus (1993) notes, has become primarily a technology for the production of scores for student performance. Writing assessment, as we mostly now know it, is a product of the search to solve the technical problem of interrater reliability. If assessment is research, then methods like constructing rubrics, training raters and the like should be secondary to the questions for which the research is being undertaken in the first place (Johanek 2000). Unfortunately, as I detail in the earlier part of this chapter and as almost anyone who has worked with assessment knows, these methods have become what most practitioners consider writing assessment itself. The result is that instead of allowing us to think about what we want to know about students, most writing assessments require extensive attention to the writing of prompts and rubrics, the training of raters, and ultimately the production of reliable scores.

Seeing assessment as research, I believe, is a way of bringing a new understanding for assessment as something that all of us who work in education might and should want to take part in, whether it be for the ethical reasons Larry Beason (2000) argues for or for more programmatic, pedagogical or theoretical reasons. Writing teachers and program administrators can recognize the necessity of asking and answering questions about their students and programs, though they might rightly resist being part of a production line that manufactures student scores according to a well-defined but arbitrary technological routine. The role of teachers also changes dramatically when we see and implement

assessment as research rather than technology. Instead of being technicians who implement a specific set of procedures, assessment as research gives teachers, administrators and other local participants the opportunity not only to control and design all aspects of the assessment, but also to build pertinent knowledge bases about their students, curriculum, teachers and programs.

In this way, the argument I make in chapter four about the need to move toward more local and site-based assessment becomes an argument about the necessity of seeing and treating writing assessment as research. It should be clear, given our discussion about technology and its influence on the development of writing assessment practices, why writing assessment ended up constructed as it is. However, once we recognize the ideological, theoretical and epistemological choices inherent in a technological approach toward writing assessment, we should also recognize that we have choices about how to construct our assessments. These choices are based upon an established literature and history of empirical research into human behavior and educational practices. In other words, writing assessment as research provides us with new opportunities to understand our students' work, our own teaching and the efficacy of our programs.

I also hope that understanding assessment as research provides an invitation for those of us who teach and administer writing programs to ask questions about the teaching of writing where we work and live and to use what we know about research into written communication to answer those questions. Discarding the technological harness that has historically controlled writing assessment can empower the people who teach and run programs to become responsible for their own assessments. This responsibility brings with it the need to know and understand acceptable research practices and to realize that theory on validity inquiry is necessary information for those of us who would conduct assessment research. For not only do we need to ask and answer questions about our students, teaching and programs, but we must also maintain that inquisitive eye on our own research practices, building arguments for the accuracy, effectiveness and the ethics of our own assessments.