# 1

# LARGE-SCALE WRITING ASSESSMENT PRACTICES AND THE INFLUENCE OF OBJECTIVITY

Assessment and objectivity have a long-standing conceptual link in the history of education in the United States. Regardless of the specific subject matter, the job of "testing"—a colloquial synonym of "assessment"—has been to arrive at some "accurate" or "truthful," i.e., objective, measurement of a student's ability. The connection between assessment and objectivity, however, is neither necessary nor absolute, in spite of both professional and popular tendencies to join the terms. The connection is, in fact, both historical and rhetorical, at least in the case of writing assessment.

The story of large-scale writing assessment in the United States is inextricably tied to the rise of positivism in the early decades of the twentieth century. During the nineteenth century, as assessment practices shifted from oral to written examinations, educators increasingly relied specifically on examinations of writing in English in large part because of persistent increases in student enrollments and Harvard's unorthodox embrace of English as the language of choice. The use of written compositions was driven in part by a desire among educators to develop effective educational methods. Evaluating those compositions, however, was another matter. Although Harvard's faculty complained most about the quality of the compositions they received from students, the substantial variation among teachers' "marks," or grades, broke the camel's back. Educators and researchers of the late nineteenth and early twentieth centuries tended to equate such variance with inefficiency and with a subjective approach to writing assessment that became increasingly indefensible and uncomfortable in the face of positivist science. In response, writing teachers developed composition scales as a method for regulating the evaluation of student compositions. At the same time, however, educational measurement obtained disciplinary status as the Stanford-Binet scale and standardized testing gained popularity and authority. These methods relied on a positivist epistemology and boasted an efficiency and objectivity that composition scales could not hope to achieve. In the face of such an obvious solution to the "problem" of excessive subjectivity,

research into the assessment of actual pieces of writing all but ceased by 1930.

This chapter analyzes this historical shift from direct to indirect assessment methods for the ways in which assessment came to incorporate an objectivist paradigm. The replacement of evaluation of student compositions (direct assessment of actual pieces of student writing) with standardized tests (indirect assessment of writing-related skills) as the instrument of choice reflects the broader social acceptance of positivist science at the time. But even as positivism's influence waned, objectivity remained as a pertinent attribute of assessment. During the early decades of the twentieth century, "reliability" and "validity" first appeared, attached to objectivist assessment as principles for measuring the qualities of any given instrument. Direct assessment methods most clearly failed the test of reliability, which measures the reproducibility of test scores. In classical educational measurement theory, reliability was—and still is—a prerequisite for validity; consequently, direct methods failed the validity test as well.

The educational measurement theory embodied in these terms continues to dominate large-scale assessment practice. In the 1960s, as composition studies developed disciplinary status, direct assessment instruments reappeared in the form of impromptu essay examinations, but researchers had to devise holistic scoring, a method that secured reliability, before institutions would accept their results. Moreover, impromptu essays gained their substantial following by arguing specifically that direct assessment methods were more *valid* than indirect ones. Much of the recent literature on portfolio assessment presents a strikingly similar justification. Even as social constructionist and postmodern theories have become mainstays in other areas of composition studies, assessment scholarship continues to rely on objectivist theoretical principles from educational measurement.

These terms—"validity" and "reliability"—have become a normalized part of assessment discourse within composition studies. Large-scale assessment of writing arose historically at the same moment as positivist science in the United States. Since positivism could successfully address the assessment concerns of the majority of writing teachers at the time, the connection was obvious and nearly inevitable. By the time writing teachers became dissatisfied with objective measures of literate ability, "validity" and "reliability" had become "normal," functioning as a terministic screen that has successfully deflected questions about the necessity of a connection between writing assessment and objectivity. This "normal" connection persists, in spite of our discipline's tendency to reject objectivist epistemologies.

## FROM ORAL TO WRITTEN EXAMINATIONS

Orality dominated the curriculum in the first half of the nineteenth century at all levels of schooling in the United States. Oral testing done by the teacher at the classroom level often involved "disputation"—not unlike the process Quintilian describes in *Institutio Oratoria*—which required that students present a thesis and defend it against other positions offered by the teacher and classmates. Its aim was also not unlike Quintilian's, as Andrea Lunsford points out: "to produce good citizens skilled in speaking, who could use the arts of discourse to influence the life of society" (1986, 4). Among its most important characteristics, disputation brought "all the language skills—reading, writing and speaking—to bear on problems of public concern," and it presented a dynamic and collaborative model of learning (1986, 4). Ideally, classroom-level testing encouraged students to make connections among various subjects and to apply that formal knowledge to situations they might confront as members of a larger society. Today, we would call this type of assessment contextually-dependent or contextually-aware.

Similarly, for promotion and graduation, students were tested either by travelling examiners or by prominent members of the community in which the students resided and would be expected to participate as citizens (Mann 1845; Williamson 1994, 154). During these examinations, examiners asked students a series of questions about their studies to which they responded orally. Because the students were tested all together and because time was limited, each student answered only a few questions and no two students received the same questions.

While these tests promoted ideals of good citizenship, they were time-consuming and became even more so as enrollments increased. Written tests represented a particularly popular response to these numbers; the first formal discussion of them appeared in 1845 with a report on the Boston Public Schools' practices in the *Common School Journal* (Mann 1845).[2] The experimental tests described in the report ask students to identify, define, and discuss particular aspects of the curriculum, including grammar and language, in the form of what we would now call short answer responses. Horace Mann,[3] editor of the *Journal*, responded enthusiastically, listing seven reasons for the superiority of written exams over oral exams:

1. they are more impartial, asking all students the same questions at the same time;
2. they are more just to the students, allowing them time to collect themselves and answer to the best of their ability;

3. they are more thorough in that they allow the examiner to ask more questions and thus test a broader range of the students' knowledge;
4. they prevent the "officious interference" of teachers who occasionally prompt students with information that will help them answer the examiners' questions;
5. they determine whether students have been taught to apply what they have learned rather than just to recite factual information, the latter of which would indicate a failing on the part of the teacher more than the student;
6. they eliminate the favoritism—both real and presumed—of the examiners; and
7. they provide a record rather than a memory or rumor of the examination—"a sort of Daguerreotype likeness, as it were, of the state and condition of the pupils' minds, is taken and carried away for general inspection." (1845, 330–34)

Mann's reasons are primarily pedagogical: written examinations demonstrate students' knowledge better than oral examinations do. Yet they also emphasize distance as a commendable quality in testing, particularly through a separation of the examiner from the examinee, and even the examinee from the examination. Written tests are better than oral examinations, according to Mann, in large part because they foster objectivity: they are impartial, neutral, and reproducible. The separation would allow anyone to see "the state and condition of the pupils' minds" at any time, including in the absence of the student. At the time of Mann's writing, and well into the twentieth century, educators and educational theorists held that knowledge could be transferred unmediated between the mind and the world outside via language; objective testing of writing regularly takes this as a major premise. Moreover, the objectivity admired by Mann was matched by the efficiency of the new tests. Under the oral system, examiners in larger public schools had five minutes or less to question each student; under the written system, all students could be tested for a full examination period (Witte, Trachsel, and Walters 1986, 16–17). If students are all asked the same questions at the same time and more questions are asked, then—at least in theory—the time and resources of both the students and the schools, not to mention the community members, are used more productively. It is interesting to note, however, that Mann's article still advocates oral examinations for those who would be teachers.

For the Boston examinations, writing was more the medium than the object of assessment, although there were sections of the test that asked

students to "parse" sentences, for example. Thirty years later, in a move so influential that it would lead to the birth of the College Board, Harvard specifically tested writing ability on a large scale when it instituted its entrance exam in English composition. But while objectivity and efficiency encouraged the adoption of written examinations in Boston, Harvard's reasons for testing writing were more political. During the second half of the nineteenth century, Harvard waged a campaign to promote the study of English over the study of Latin and Greek, which until then had been the hallmarks of proper education at both the secondary and post-secondary levels. Under the guidance of Charles W. Eliot, president from 1869 to 1909, Harvard added modern languages and literature, as well as experimental sciences, to the curriculum (Kitzhaber 1953, 28–29). Under Adams Sherman Hill, Boylston Professor of Rhetoric appointed by Eliot, Harvard's requirements in written composition increased, becoming by 1879 the only requirement after the first year of college (Hill, Briggs, and Hurlbut 1896, 16, Appendix). Both Eliot and Hill firmly believed that in an age of positivist science, the study of classical languages was nearly useless and certainly outdated when compared to the study of English, a tool for everyday use (Kitzhaber 1953, 28–29, 53–9). One expression of their views required all candidates for admission to pass an entrance examination in English: the first of these appeared in 1865–66 as a requirement for reading aloud; the written exam itself dates from 1874–75 (Hill, Briggs, and Hurlbut 1896, Appendix).

Unfortunately, the natives could not write the language to Harvard's satisfaction. In a series of essays written between 1879 and 1892 and distributed as a pamphlet in 1896, Hill, L.B.R. Briggs, and Byron Hurlbut complain about the candidates' inability to write grammatically correct and rhetorically elegant prose. Hill's essay, for example, catalogs the errors in spelling, mechanics and grammar in the examination books of June 1879, concluding that "[m]any books were deformed by grossly ungrammatical or profoundly obscure sentences, and some by absolute illiteracy" (1896, 10). Of the 316 examinees, only 14 passed "with credit" (1896, 9–11). Hill's primary complaint was that Harvard had to fix these errors. He argues that Harvard, "the university which professes to set up the highest standard in America," should be given better material from which to make "educated men" (1896, 11). All three authors in this pamphlet blame the secondary schools. In response, the secondary schools tried to prepare students for college entrance examinations, whether those students were planning on attending college or not, to the exclusion of more general instruction in writing. The result was that many

students could write an essay on the texts they had studied but were nearly incapable of writing on any other subject (Kitzhaber 1953, 71).

The development and application of Harvard's entrance examination in English had far-reaching political and pedagogical consequences. Unlike Mann, Harvard's leaders did not find merit in the objectivity or efficiency of the exam; they saw it as a necessary means of maintaining the university's elite position and mission in higher education. Under a banner of standards, Harvard banished instruction in grammar to the secondary schools and raised its admission standards in an attempt to force the lower schools to do a better job of teaching writing, specifically in English. In effect, this response established the testing of writing as a gatekeeping mechanism: those who could not write well did not belong.

Moreover, Harvard's examiners defined writing well as writing correctly. The 1873–74 catalog, for example, required "a short English Composition, correct in spelling, punctuation, grammar, and expression"—the last of these meaning diction and word choice more than the creativity or style we might read in the term today. By 1895–96, the only change to these four was the rather disappointing substitution of "division into paragraphs" for "expression" (Hill, Briggs, and Hurlbut 1896, 6, Appendix). This standard established not only a legacy of grammatical obsession, but also the measure of grammatical and mechanical competency necessary to attend college. In the 1896 pamphlet, Hurlbut explains that the content of the June 1891 tests was reasonably intelligent and "well proportioned" (1896, 44). However, the applicants' punctuation and diction were unacceptable, and their grammar relied too heavily on Latin constructions.

Harvard's practices had a number of long-term consequences. While the emphasis on English composition certainly helped establish this field as an appropriate scholarly discipline, the demands for "correct" composition that Harvard educators placed on secondary schools heralded the low status of post-secondary composition instruction to-date. More importantly for this study, their grammatical and mechanical emphasis set the stage for positivist assessment methods. According to their writings, these highly influential educators stressed spelling, punctuation and grammar, aspects of writing which contemporary compositionists would consider rudimentary and mechanistic. The most advanced or rhetorically complex element—"expression"—became merely a requirement for paragraphing in the later examinations. Thirty years later, the College Board combined objective testing methods with these standards, so that tests of English composition became tests of discrete grammatical skills and required no writing at all.

Surveying societal trends during these early years, Lunsford concludes that the pressures behind the decline of oral examinations and the rise of written ones came from burgeoning college and university enrollments, the growing influence of scientific methodologies, and the increasing emphasis on writing in colleges (1986, 5). Of these reasons, the influence of science—and more specifically, the approbation for objectivity—was the most far reaching in the history of writing assessment. While increasing enrollments at all levels bolstered demands for more efficient testing, the rising numbers could just as easily have led to calls for more teachers and schools, or even less testing. However, positivism, a flourishing endeavor in the nineteenth and early twentieth centuries, argues the laws of nature are knowable and applicable through science and relies on measurement as a primary tool in the search for these principles. Mann's "Daguerreotype likeness," for example, reflects a positivist belief that the faculties of the mind can be readily seen in writing produced by that mind. The increased interest in writing coming from colleges following Harvard's lead meant that the pictures most often took the form of compositions; the cultural affinity for objectivity helped establish the criteria by which those pictures would be judged.

## COMPOSITION SCALES

The institutionalized testing of writing begun by the Boston schools and the Harvard exams brought the interests of writing and science together in writing assessment by the end of the nineteenth century in such a way as to push efforts at standardization to the fore. Researchers found that as teachers and institutions evaluated writing, their standards for assessment varied significantly from examiner to examiner and even from scoring session to scoring session for the same examiner.[4] Teachers, administrators, parents, and college educators equated this irregularity with a lack of standards, a condition intolerable in an age of scientific education, and their complaints about this unreliability spurred attempts by educational theorists and researchers to find ways to measure writing ability with scientific precision. The "solution" they developed in the early twentieth century was a series of what were called "composition scales."

Composition scales consisted of a graduated series of essays that served as models for comparison with actual student compositions to determine the merit of the students' work. The first of these was developed for use with secondary school students by Milo B. Hillegas, a professor at Columbia University, and was published in 1912. From 7,000 compositions originally collected, Hillegas chose 21 papers for his scale, mostly

written by secondary school students, but supplemented on the high and low ends by compositions produced by adults or taken from the early writings of literary authors because Hillegas could not find actual student writing that would fill out the scale. Ideally, anyone who wished to determine the merit of a student composition could hold it up to the samples comprising the scale, find the nearest matching model, and deliver a "correct" score. Hillegas's scale did not attempt to differentiate among grade levels, although later supplemental scales and adaptations did.

While scholars, administrators, and teachers writing at the time criticized Hillegas's scale in its particulars, many of them enthusiastically embraced the idea of uniform standards.[5] These scales at least theoretically addressed a pressing need in education for some sense of agreement about grading, scoring, and marking. Even those who disagreed with the scales in principle admitted that the inability to deliver consistent scores on compositions indicated a deplorable absence of standards. Thomas H. Briggs (1922), for example, finds the scales limited, but he still argues that schools, teachers, and departments should devise scales of their own to determine at least grade level promotion, and that any scale used intelligently would be an improvement. He points out that "the alternative seems to be a frank admission that English teachers can not discriminate qualities of composition with sufficient accuracy to gain credence in the reliability of their marks, acceptance of the principle that all pupils shall have the same number of semesters of instruction, each one profiting as he may and no one failing if he is earnest in effort" (1922, 442). With positivist science gaining popularity and influence throughout American society, teachers' marks needed to be consistent in order to gain even a modicum of respect from those outside of education, including those who funded the schools. This condition reflects the value of objective measures at the time; without reproducible results, teachers' marks provided no rational justification for assessment.

Hillegas (1912), Frank Washington Ballou (1914), M.R. Trabue (1917), and others who developed and applied these scales saw the need for uniform standards as their primary purpose. Standards, Hillegas argues, would lend credibility to education: "If there were standards or scales for the measurements of results in the various school subjects that would approximate the accuracy of the scales used in measuring extension, weight and time, educational administrators and investigators would be able to measure and express the efficiency of a school system in terms that would carry conviction (1912, 2). Hillegas's concern for standards that could "carry conviction" derives in part from a desire to apply positivist

theories to educational practices. In positivist science, what is real can be observed the same way by multiple observers. To proclaim itself scientific, composition scoring would have to become, at the very least, consistent; if such testing could prove itself scientific, it would be beyond reproach. In this manner, researchers such as Hillegas became invested in an objectivist paradigm.

Consistency in marking was not the only stumbling block on the way to the laboratory. Even more problematic was the inherent subjectivity of the object of assessment. What Hillegas's scale measures—and what Trabue's and Ballou's accept as the object to be measured—is something he calls "merit" (1912, 9). In defining this term, however, Hillegas uses circular logic: "The term as here used means just that quality which competent persons commonly consider as merit, and the scale measures just this quality" (1912, 9). His definition gets no more specific than this, though he identifies "competent persons" as teachers of English at the secondary and post-secondary levels, literary authors, and psychologists "familiar with the significance of scales and zero points in the case of intellectual abilities and products" (1912, 12). In effect and reminiscent of many early forays into criteria for assessment, Hillegas claims to measure merit without defining it except to insist that those who know it will recognize it. In so doing, he works into the equation a subjective judgment not unlike teachers' marks.[6]

The subjectivity of the judgments required by those developing and applying the scales was largely ignored; the inconsistency, however, was not. Ultimately, these scales resulted in scores as inconsistent as teachers' marks without the benefit of scales (Kelly 1914; Sackett 1917). Without consistency, the scales could not claim uniform standards; without uniform standards, they could not claim objectivity; and without objectivity in an age of science, they had to be replaced.

## INDIRECT TESTING OF WRITING

During the first three decades of the twentieth century, substantial effort went into refining methods for assessing student writing, but by the late 1920s most research into the direct assessment of writing—assessing of actual pieces of writing—including composition scales, had stopped. R.L. Lyman's *Summary of Investigations Relating to Grammar, Language and Composition* (1929) summarizes the primary research findings from the turn of the century through its publication, and there is considerable work on direct assessment. With some exceptions, particularly in British journals (e.g., Cast 1939a, 1939b; Vernon and Millican 1954), there is

little on direct assessment after this until the holistic movement almost 40 years later.[7] As Norbert Elliot, Maximino Plata, and Paul Zelhart (1990, 30) point out, at the time the methods for direct assessment were becoming more standardized—as societal norms required—but they were still time-consuming and reliant on subjective judgments except where tests were limited to grammar and mechanics. Research might have been more vigorous but for the development of indirect, "objective" assessment methods coming out of educational psychology. The efficiency of these "new-type" tests far exceeded even the most optimistic expectations for direct assessment. There was no contest.

The first developments in indirect testing came from France at the turn of the century when Alfred Binet and his colleagues developed tests to determine the abilities of school children, particularly those in need of additional or special assistance. Like Hillegas's measurement of composition, Binet's method for determining the mental age of a child—and thus determining whether or not the child was "subnormal"—was a scale, in this case, a series of thirty tasks of increasing difficulty (Chapman 1988, 19–20). Binet's first scale appeared in 1905; he followed it with a revision designed for testing all children in 1908 and a second revision in 1911. Lewis M. Terman, a researcher at Stanford, modified it for use in the United States in 1916 and renamed it the Stanford-Binet.

Both the Binet scale and the Stanford-Binet were labor-intensive measurements conducted through interviews and observations which required "the time and efforts of one examiner working with a single student for a period typically involving several hours, including the tasks of test administration, test scoring, and evaluation of results" (Williamson 1994, 156). Moreover, the interviewer was crucial to the process and was expected to intervene during the testing to "probe ambiguous answers" (Williamson 1993, 23). In an exercise of American ingenuity, Terman's student, Arthur Otis, developed a multiple-choice version of the exam that expedited the process and erased the interviewer. Terman adapted this test for the United States Army to sort recruits for officer candidacy and specialized training during World War I. The American adaptation eliminated the labor-intensive aspects of the Binet scales while preserving its use as a sorting mechanism.

Shortly following World War I, Terman and a group of his colleagues received funding from the National Education Association to adapt the "group testing" developed for the war effort for the purpose of reorganizing the public schools (Terman et al. 1922, 2). The reorganization resembled what we might call "tracking," slating some secondary students

for vocational coursework and others for college preparation. This tracking institutionalized self-fulfilling prophecies about whether the students would be attending college: those on the academic track generally went to college, while those on the vocational track usually took up a trade after high school. Michael Williamson argues that the nation's need for organizing its fighting force provided the jump-start for objective testing, but that the rise of mass public schooling supported its longevity (1993, 24). It did not hurt that both conservatives and liberals agreed on the value of objective tests, although for different reasons. Conservatives approved of them because they reinforced uniform standards for certification in the subject areas and encouraged discipline. Liberals found them useful to diagnose student needs so that teachers could address them on an individual basis (Applebee 1974, 95).

The College Board, originally developed to standardize college admissions testing, also began using these "new-type" examinations and took objective testing to new heights. Prior to 1894, each college developed its own entrance examination with its own required reading list. In an attempt to standardize the lists, the National Conference on Uniform Entrance Requirements, a joint effort among east coast professional educational associations, was formed in 1894, and the College Entrance Examination Board (CEEB)—which came to be known as the College Board—grew out of that original conference. When the CEEB administered its first examination in 1901, it used the Restricted Examination in English—an essay test drawn from a list of literary texts. The early Restricted Examinations required "a good memory for textual features such as plot construction and descriptive details" (Trachsel 1992, 76). In 1916, in response to the complaints of secondary teachers about the emphasis on textual detail, the College Board introduced the first Comprehensive Examinations in English which would test the examinees' abilities to interpret texts more generally and to express themselves in writing (Trachsel 1992, 77–79; Elliot, Plata, and Zelhart 1990, 31).[8]

In 1921, Edward L. Thorndike, a colleague of Hillegas, demonstrated that the objective "mental" tests from the field of psychology were better predictors of college performance than the College Board's existing essay entrance examinations or the student's high school record (Trachsel 1992, 108–9). Based on this and other studies, the College Board administered its first Scholastic Aptitude Test (SAT) in 1926. Offered in a predominantly multiple-choice format, the SAT quickly became popular for its efficiency, particularly with the governing bodies of universities—bodies that were and are increasingly made up of corporate and financial

executives (Trachsel 1992, 108–09). The first SAT exams took an analytic approach to reading and writing skills, dividing each into a series of discrete skills to be tested individually—an approach commensurate with the technocratic model of literacy prevalent at the time, which I will discuss in chapter two.

Not everyone, however, welcomed the SAT with open arms, even in the ranks of the College Board. In 1929, the Board appointed a Commission on English to determine the effectiveness of the SAT and the Restricted and Comprehensive Examinations in English. The Commission did not approve of the SAT, but they did find that the results of the SAT correlated well with a student's ability to succeed in first year English. The Comprehensive Exam—which the Commission favored—did not. These findings were buried in Appendix D of their 1931 report (Commission on English 1931, 261–74). The Commission argued that objective tests were of limited use for measuring "creative values" (1931, 210), i.e., those reflecting students' ability to summarize and interpret passages of literature (1931, 154). This emphasis on literary skills led to the retention of the composition portion of the English entrance exam, at least in the short run, but the questions moved away from general experience and toward more poetic and literary topics (Trachsel 1992, 116–21). In 1934, however, in the face of the effectiveness and popularity of the SAT, the College Board discontinued the Restricted Examination in English, and in 1942, the Comprehensive Exam (Trachsel 1992, 110–11; Elliot, Plata, and Zelhart 1990, 34).

Williamson argues that a combination of factors led to the rise of large-scale objective testing. These factors included a belief in meritocratic ideals, the promise of positivist science, questions about the fairness of examinations—particularly the variability of subjective scoring—and calls for accounting of the funds spent on education. In comparison with essay exams and other labor-intensive efforts, large-scale objective testing provided what seemed to be a fair and consistent way to gather information about students. Moreover, because they were standardized, the test results could be used to compare schools and provide information to agencies, parents, and governing bodies (Williamson 1994, 159–60). These "benefits," however, also impeded curricular development at the elementary and secondary levels. Trachsel points out that because the tests were so important, teachers had to teach to them, and even when crises called for significant reform of the curriculum, the predictive nature of the tests tended to reinforce the status quo (1992, 172). Concerns such as these, however, were insignificant at the time when compared to the benefits

accrued by objective testing. Objectivity provided strong answers to the most pressing questions of the day; the fit between assessment and objective principles seemed only natural.

## RELIABILITY AND VALIDITY

In 1922, Thomas Briggs used the term "reliability" in connection with the variable results of testing done by individual teachers (442). This term and its partner, "validity," were derived from psychological testing, specifically psychometrics, a branch of psychological testing using quantitative methods to measure mental qualities such as intelligence. Psychometrics emerged from the positivist belief that anything that exists is measurable, and the corollary that mental processes exist and therefore are measurable. In writing assessment, the application of psychometric principles also depends on the premise that written abilities mirror mental capabilities. Prior to World War I, the terms "reliability" and "validity" had no particular currency in educational assessment, and no one used them. Since World War I, however, these terms, which signify objective assessment principles, have governed writing assessment scholarship and practice to the exclusion of alternative theories.

Reliability carried more weight than validity for most of the twentieth century. Reliability "refers to the *reproduceability* [*sic*] of a set of test results" (Lyman 1991, 22). In practice, a test cannot be considered reliable unless it consistently provides the same results or nearly the same results, regardless of the conditions under which it is scored. The multiple-choice format of the SATs claims high reliability because the machines that score it do not wake up on the wrong side of the bed or become enamored with a clever turn of phrase. Essay tests are necessarily less reliable.

Reliability has remained an issue throughout the history of large-scale writing assessment in the United States. The first written tests were developed not only to meet the need to test large student populations, but also to address concerns about fairness to students and favoritism by teachers. In oral examinations, each student was asked a different question, so because there was necessarily some variation in the level of difficulty of the questions asked, this testing was ultimately uneven. In contrast, written assessment asked all students the same questions at the same time with less teacher intervention. Moreover, their responses could be compared, not only within classes, but also across schools, districts, and states.

The comparison of results in situations such as college admissions testing led to concerns for ways to set and measure standards so that comparisons would be meaningful and fair to the schools, students, and

teachers involved. Hillegas takes comparative standards as the primary motivation for his scale:

> [e]very attempt to measure the efficiency of instruction in a school or system or to evaluate different methods of educational procedure serves to emphasize the importance of standards. Proper standards would make it possible to compare with certainty the work done in one school or system of schools with that done elsewhere. They would make it more difficult for mere opinion to control so much of our school-room practice. (1912, 1)

Both Hillegas (1912) and Trabue (1917) went to great lengths to assure the reliability of their samples. For each scale, they sent their samples to multiple readers, and they only chose the samples whose scores demonstrated the highest reader agreement. In addition, Hillegas spends a fair amount of time in his essay explaining the normal deviation from the "true" score of each sample—that such deviation is regular: for every high score there will be an equally low score, in effect canceling the difference. From a contemporary perspective, his explanations sound somewhat defensive and contrived. His science, however, differs from our own, and his explanations conform to the expectations of his day. Our own embrace of deviation would likely sound lackadaisical, at the very least, to his ears.

In spite of efforts such as Hillegas's, reliability problems contributed to the downfall of those early direct assessments of writing. As Elliot, Plata, and Zelhart point out, the inter-rater reliability among the original readers was so low in the Hillegas scale of 1912 that only 21 papers out of 7,000 samples collected generated sufficient agreement to appear in the published version; in the Nassau County supplement of 1917, only 30 out of 5,500 made the final cut (1990, 29). They argue that research in direct writing assessment was thriving during the first thirty years of the twentieth century, but that "the methods of evaluation were extraordinarily time consuming, yielding low rates of inter-reader agreement. Moreover, the claims made for the scales were greatly exaggerated" (1990, 30). Although direct writing assessment might have improved, indirect assessment was far more efficient, and the combination of "the rise of the College Board, the beginning of the efficiency era in education, and the growth of intelligence testing" undermined work in the direct assessment of writing (1990, 35).

Edward M. White, perhaps the best-known contemporary writing assessment scholar, calls this quality "fairness." In translating "reliability" into "fairness," however, White oversimplifies the term and conflates ideals of

consistency, objectivity, and ethics. In psychometrics, reliability is a technical measure of consistency; White's implicit argument is that in order to be ethical, evaluators must be technically consistent, and in White's configuration, consistency depends on objectivity. This is the reasoning used by most contemporary compositionists who favor the application of psychometric principles: in order to be "fair," evaluators must be objective. This equation makes a certain sense, but its logic is loaded: who among us would want to be "unfair" to students? Refuting this claim requires that the responder carefully disentangle ethics, objectivity, and consistency—a substantial task and one that few assessment scholars in composition have even attempted. Some contemporary scholars challenge the need for reliability by treating it as irrelevant or inconsequential, but most have not refuted this logic directly. Consequently, even though its usage is imprecise in composition studies, "reliability" remains part of the writing assessment landscape as a means of assuring that we treat our students justly when we evaluate their work.

Validity, reliability's partner term, refers to the ability of the test to measure what it is required to measure, its suitability to the task at hand.[9] In classical testing theory, the dominant approach through most of the twentieth century, a test is valid if the results of that test have some determinable connection to the actual competencies, aptitudes, skills, or knowledge the test purports to measure. For example, a twelve-inch string is both valid and reliable for measuring one foot; a thirteen-inch string is reliable in that its length does not vary, but it is not valid for measuring one foot. Tests that measure mental skills are hardly as simple, but the principles are basically the same.

Multiple-choice and other "objective" tests originally proved their reliability through the use of questions with ostensibly only one correct answer, which eventually machines could score. Their validity seemed to come as part of the package, except that the dictates of reliability meant that only what could be agreed upon as "correct" could be tested. The result: grammar, mechanics, usage, vocabulary, and the like dominated objective tests of writing ability. This need for unequivocal "correctness" coupled with Harvard's endorsement of these standards encouraged test designers to rely on this narrow definition of "writing."

Our contemporary disapproval of this oversimplification, however, does not mean that no correlation exists between the results of such tests and other "measures" of writing ability. As Roberta Camp points out: "From the perspective of traditional psychometrics, in which high test reliability is a prerequisite for validity, the multiple-choice writing test has

also been seen as a valid measure. 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" (1993, 47). The 1931 Report from the Commission on English produced the same results, and this correlation substantiated much of the use of objective tests during the twentieth century. In effect, psychometrics claims that a test cannot be valid if it is not reliable, and reliability has been defined in terms of consistency of scoring, which is why evaluators have tended to prefer machine-scorable tests. Given the cultural affinity for efficiency and scientific precision of the early twentieth century, it is hardly surprising that educators and scholars used these principles to bolster calls for objective tests at the time.

More importantly, these principles gathered significant currency during this early period, so much so that they have influenced large-scale writing assessment to the present day. There was no effective competition, so the terms became dominant and then normal: all assessments were, as a matter of course, subject to the criteria of validity and reliability. Contemporary compositionists continue to invoke these terms to substantiate their assessment procedures, often without questioning their applicability to the evaluation of writing and even assuming a necessary connection between these principles and any assessment procedure, regardless of subject matter. Through these normalized principles, even contemporary writing assessment practices carry an objective orientation.

## HOLISTIC SCORING

While multiple-choice remained the test format of choice, between 1954 and 1971 the College Board made a series of concessions to educators who insisted on the direct assessment of writing. From 1954–56, for example, the board offered the General Composition Test, a two-hour, impromptu exam consisting of one question on a popular topic, as an alternative to the objective English Composition Test. These essay exams were scored by trained readers according to predetermined criteria—in this case and in order, mechanics, style, organization, reasoning, and content—a process known as analytic scoring (Trachsel 1992, 148–49). Such innovations, however, did not last long during this period.

Ironically, test development specialists at the Educational Testing Service (ETS) are the ones who devised holistic scoring in response to their smaller clients who wanted to see actual pieces of writing and who

were less concerned with efficiency. Working for ETS, Paul B. Diederich, John W. French, and Sydell T. Carlton (1961) determined that the lack of inter-reader reliability in scoring essays arose from differences in the criteria for judging essays. In their NCTE publication, Richard Braddock, Richard Lloyd-Jones, and Lowell Schoer (1963) took this idea a step further and argued that readers must help develop the criteria to be applied and must review and apply these criteria periodically to ensure that they continued to agree. Fred I. Godshalk, Frances Swineford, and William E. Coffman (1966)—also researchers at ETS—took up this research and published a monograph on holistic scoring. They solved the reliability problem of scoring compositions by limiting the number of topics on which students would write and by training readers to develop and discern particular criteria.

This last process is an early version of the procedure that composition programs adopted during the 1970s and 1980s. Godshalk, Swineford, and Coffman devised holistic reading of essays—which they define as a scoring method in which "readers [are] asked to make a single judgment with little or no guidance as to detailed standards"—to address the dual problems of "reading reliability and the burden of a slow analytical reading" associated with essay scoring to-date (1966, 1–2). They solved the "reading reliability" problem by asking readers to score a series of carefully chosen sample essays (on a 3–point scale in this case) and then to publicly compare the scores. The comparison of scores—what we now call "norming"—had the controlling effect on the readers it was intended to: The researchers reported that "[n]o effort was made to identify any reader whose standards were out of line, because that fact would be known to him [*sic*] and would be assumed to have a corrective effect" (1966, 10). Upon finding that they had obtained statistically sound reliability, the team reported the experiment a success.

From our current-day perspective, however, the success was at best partial. True, holistic scoring helped locate writing more centrally in writing assessment. But nowhere in the Godshalk report does the reader get a sense that writing has any particular value as a whole, as an activity, as a method of instruction. Instead, the authors tend to treat writing assessment as a puzzle to be worked out. Or, in the terms I have been developing thus far, as a problem of positivist science: writing exists; therefore, it can be measured, and we can truly know it only through that measurement. The trick was to find the right yardstick. In fact, the introduction to the Godshalk report—written by Edward S. Noyes, special consultant to the president of the CEEB and a College Board researcher during the

1940s—declares "that this problem [the measurement of a student's ability to write] has at long last been solved" (1966, iv).

White argues that ETS developed holistic scoring "as a bone to throw to English teachers" while the multiple-choice tests provided the actual data for testing experts, and that the supporters of holistic scoring in ETS were and are few (1993, 82). The report by Godshalk, Swineford, and Coffman bears this out at least in part: "In spite of the growing evidence that the objective and semi-objective English composition questions were valid, teachers and administrators in schools and colleges kept insisting that candidates for admission to college ought to be required to demonstrate their writing skill directly" (1966, 3). While the researchers do not entirely ignore these pressures, they do make it clear that for the purposes of *their* testing, the essay exam adds nothing; only in pedagogical eyes are direct measures of writing necessary. They return to this idea in their conclusion, where they argue that the best test in statistical terms combines the objective and direct writing aspects (1966, 41). However, they point out that if cost is factored into the statistical equation, and if cost carries any real import, then direct assessment measures alone—which are substantially more expensive to score than the objective measures—are clearly *not* any better than the objective measures alone. They find the value of direct assessment elsewhere, arguing that "the advantage [to evaluating actual pieces of writing] has to be assessed in terms of the model the essay provides for students and teachers" (1966, 41), but the authors are not clear on what the pedagogical uses of essays might be.

White, however, tells the pedagogical side of the tale in "Holistic Scoring: Past Triumphs, Future Challenges" (1993).[10] He suggests that holistic scoring is, in a sense, a product of its time. It emerged during a period when educators and students began challenging the privilege of "correct" English, when poststructuralism, writing research and writing scholars appeared, when universities opened their doors to non-traditional students, and in the wake of the student rebellions of the 1960s (1993, 83). He describes the mood of those scholars and teachers working with holistic scoring:

> Those of us who were involved in the missionary activity of promulgating holistic scoring of student essays in the 1970s tended to feel that we had achieved the answer to the testing of writing. By developing careful essay questions, administering and scoring them under controlled conditions, and recording a single accurate score for the quality of writing as a whole (with scoring guides

> and sample papers defining quality), we had become committed to a flexible, accurate, and responsive measurement method, one that could come under the control of teachers. (1993, 79)

The benefits were many. On the social front, holistic scoring addressed the inequities of class by providing "a procedure that defined writing as producing a text, that could award scores for originality, or creativity, or intelligence, or organization, as well as mechanical correctness in the school dialect" (1993, 86). In contrast, objective tests of writing ability, such as the Test of Standard Written English, tally the responses to multiple-choice questions on grammar, usage and vocabulary—the grammar, usage and vocabulary of the upper middle-class white family in America—and thus define scholastic aptitude in terms of socioeconomic status. On the pedagogical front, holistic scoring brought evaluation into the classroom. Many teachers made it part of their pedagogy, including scoring guides and peer evaluation as part of their classroom materials and activities. Moreover, questionable as this claim may seem, holistic scoring brought revision into the classroom, according to White, since it made standards public and treated them as goals to be met through a focus on the writing process (1993, 89).

The sense of gain associated with the rise of holistic scoring, the success which led White to describe the movement as a "remarkable triumph" (1993, 80), was not unanimous. Some teachers found it "unsettling" because it undermined the time-honored tradition of red ink. Testing professionals often treated it as a poetic attempt by subjectivists to overcome the hard numbers of objective tests. And administrators who used it had to fight regular battles over cost and reliability (1993, 82). However, the work done on holistic scoring during the 1960s and 1970s made possible broader discussions about the relationship between assessment and the context in which the assessment takes place, discussions which had not been welcomed in the days of composition scales and intelligence testing.

Holistic scoring, however, was unable to escape the influence of an objectivist paradigm. White's celebratory account of holistic scoring, for example, codes the limitations of this process in terms of "validity" and "reliability," to the point of using subsections so titled. Under the section titled "Validity," White argues that holistic scoring relies on face validity: holistic tests that measure writing look at actual pieces of writing to do so. While face validity has a great deal of currency, at least with teachers, White observes that the "reality" of the holistic writing situation is

based on a constricted view of writing: without revision (a point which contradicts his earlier claim in the same essay), with a grader for an audience, and without purpose beyond the test question. Ultimately, though, White defends holistic scoring by arguing that the artificial situation of a holistically scored essay test beats the artifice of a multiple-choice test. Moreover, he claims that just because holistic scoring is a good idea does not mean that it is good for all situations; its uses should be limited to situations that call for general assessments of overall writing ability and the scores must be used responsibly (1993, 90–93).

In his accounts of holistic scoring, White spends almost twice as much time on reliability as validity. He points out that some areas of reliability, such as variability in student health on test days, are beyond the realm of test development and influence all tests. He argues, however, that those elements of reliability within the test developers' influence should be addressed. He outlines potential problems in the misdevelopment of test questions and the mismanagement of reading sessions, and he emphasizes the increased cost of holistically scored essays. He argues that readers should be encouraged to develop a sense of community through socializing so that they will be more willing to cooperate and see each other's points-of-view. His emphasis on reliability functions as a warning to teachers and administrators: follow these procedures carefully or the testing experts will take your essays away from you—an argument that White has made repeatedly in his career.

The acceptance of holistic scoring marked the return to the direct assessment of writing. Unlike earlier scholars working on direct measures such as composition scales, researchers working in holistic scoring focused their energy specifically on proving its validity and reliability. Testing experts and administrators, then and now, tend to favor indirect assessment, arguing that multiple-choice examinations allow for high reliability, reasonable validity, and relatively efficient administration and data collection. Composition scholars, among others, favor direct assessment of writing, arguing that it is more valid because it measures writing ability by examining actual writing, that it can be almost as reliable as indirect assessment, and that the costs are roughly equivalent considering that indirect assessment has as large a pre-test cost (for continuous design) as direct has a post-test cost (for scoring). Composition scholars also favor direct assessment because it encourages aspects of writing that educators value, including process and rhetorical awareness.

Validity and reliability have governed the scene of large-scale writing assessment through most of the twentieth century and into the current

one, determining in advance the standards to which all assessment methods must measure up. Much of the literature debates the relative value of these principles but remains nonetheless locked into addressing them. Early in their history, indirect writing assessment methods employed these terms, as part of the package of psychometrics, to bolster their legitimacy. In claiming these principles, indirect assessment accrued the advantage of defining the terms of the struggle over the legitimacy of any testing measure. Holistic scoring fought its primary battle proving that it could be as reliable as indirect assessment methods. The question was always whether or not holistic scoring could measure up. Holistic assessment has "face validity" (it looks like the "right" thing to evaluate), but with the emphasis in its literature on inter-reader reliability, its aim is ideally to be as objective as multiple-choice examinations. Essay tests were discredited in the 1910s and 1920s precisely because of problems with reliability. They reappeared when they could be scored like objective tests, and at the time, no one much questioned the appropriateness of objective criteria.

## PORTFOLIO ASSESSMENT

Used for many years by the fine arts, portfolios are a way of collecting materials, rather than an actual evaluation process. In writing instruction, portfolios became popular during the early 1990s as an extension of the single-sitting impromptu essay writing assignment that resulted from the development of holistic scoring. While the impromptus looked more like "real" writing than did multiple-choice questions, they still did not match the process pedagogy employed in many, if not most, composition classrooms. Specifically, impromptu exams had no provisions for revision and did not allow students to demonstrate a range of abilities, both of which portfolios could do.

The movement for portfolio use came from within the ranks of writing teachers—not since composition scales had this been the case—but like holistic scoring, the College Board experimented with the idea first. In the early 1980s, ETS developed a proposed Portfolio Assessment Plan in response to educators' demands that testing reflect pedagogical theory. The plan would have asked students to submit a variety of essays, including a student-selected piece and an introductory letter addressed to admissions officials or prospective employers. However, the plan "was abandoned by ETS on the basis of its failure to meet the agency's required standards of time- and cost-efficiency, scoring reliability, and the appearance of scientific objectivity" (Trachsel 1992, 175–76). Trachsel points out that such reasoning becomes increasingly acceptable in the face of

diminishing budgets and increasing demands for accountability (1992, 176), both trends which we are facing currently as we did in the 1980s.

Used in the context of the classroom or for program-wide placement or competency assessment, however, portfolios have become popular as locally driven assessments of writing. Perhaps because of the grassroots origins of the method—a large-scale writing assessment procedure that actually comes from teachers—researchers in portfolio assessment have resisted providing experimental and objective "proof" that portfolios are quantitatively "better" than holistically scored impromptus or objective tests. As the editors of *New Directions in Portfolio Assessment* point out, "the research on portfolios has been more classroom-based, more reflective, and more qualitative in nature" (Black et al. 1994b, 2).

Even proponents of portfolios, however, speak in ETS's terms: validity and reliability. In fact, their central claim about the value of portfolios is that they are *more* valid than holistically scored impromptus, which are *more* valid than objective tests. Peter Elbow, one of the earliest portfolio proponents, relies on these principles to make his point, even as his essays spill over with commentary about portfolios' pedagogical value. For example, in the Foreword to *Portfolios: Process and Product*, Elbow outlines reasons why portfolios hold more promise than impromptu examinations. His first reason is "improved *validity*": he argues that "portfolios give a *better picture of students' writing abilities*" (1991, xi). His second reason seems fairly perverse, but still stays within the realm of accepted discourse: he argues that portfolios are promising precisely because they complicate reliability. His point is that real reading, even by trained readers such as English professors, necessarily involves disagreement and that portfolios encourage something more like this real world reading. However, Elbow sees validity and reliability in a sort of inescapable binary that must be addressed: "Given the tension between validity and reliability—the trade-off between getting good pictures of what we are trying to test and good agreement among interpreters of those pictures—it makes most sense to put our chips on validity and allow reliability to suffer" (1991, xiii). Taking up this comment, White protests that if compositionists treat reliability and validity as oppositional and ignore the former, we risk becoming irrelevant to the larger testing community, including ETS and their allies—governmental agencies and administrations (1994c, 292). The composition community, he argues, cannot afford to be so shortsighted. Elbow argues that we have sufficient power to dictate writing assessment practices, but even as he says this, he relies on the terms garnered by indirect assessment to make his point.

In White's estimation, validity is not a problem for portfolios—they have a better claim on "real" writing than impromptus do—but reliability is. The variable contents of portfolios, even under the most directive guidelines, make it even more difficult for examiners to generate reliable scores than do impromptu essays. Multiple pieces of writing developed and revised over the period of a semester or more are harder to agree on than essays written in a limited time period in response to a limited number of prompts. White argues that "reliability should not become the obsession for portfolio evaluation that it became for essay testing," but he points out that "portfolios cannot become a serious means of measurement without demonstrable reliability" (1993, 105).[11] White clearly understands the power of the lexicon.

From the outset, portfolios have contained the potential to displace reliability and validity as the central principles in writing assessment, but they have yet to do so. The terms are, admittedly, difficult to push aside, and it is hardly wise to ignore them, considering their currency in national-level assessment practices. Collections about portfolios such as *New Directions in Portfolio Assessment* (Black et al. 1994b), *Portfolios: Process and Product* (Belanoff and Dickson 1991), and *Situating Portfolios: Four Perspectives* (Yancey and Weiser 1997) outline alternative assessment paths, but tend not to develop a research agenda that would realign the existing power structures. Most of the essays in these collections seem to have difficulty theorizing assessment, opting instead for personal narratives and local descriptions that circumvent the terms altogether.[12] While these narratives and descriptions suggest alternative assessment practices and could be extrapolated to theoretical principles, their authors seem to have some difficulty moving beyond the context of their local programs. However, the focus on the instrument—portfolios—may limit the discussion, and in chapter four, I will look at more specifically theoretical endeavors in the past decade that are not attached specifically to portfolios.

Liz Hamp-Lyons and William Condon, by contrast, specifically undertake the development of a theory for portfolio assessment in *Assessing the Portfolio: Principles for Practice, Theory, and Research* (2000). Their project is, in large part, to develop "a credible, well-articulated theoretical base" for portfolio assessment (2000, 116), and while their work is careful and well thought-out, it remains within the educational measurement tradition. Specifically, Hamp-Lyons and Condon point out, echoing prior scholars, that "before [portfolios] could be taken seriously, the issue of reliability had to be solved" (2000, xv). Now that portfolios can demonstrate reliability, they claim, "the next stage, the process of theory-making and research

that will establish portfolio-based writing assessment as firmly as, today, we see with timed writing tests" can begin (2000, xv). That is, Hamp-Lyons and Condon accept reliability as a prerequisite for theoretical work in portfolio assessment, rather than as a theoretical concept in and of itself.

The lack of alternative principles leaves portfolios subject to the dictates of reliability and validity. As such, portfolio assessment in large-scale situations often demands increased norming so that readers can learn to read the same way. It may also demand restrictions on the contents, which in turn raise questions about the pedagogical value of an assessment instrument that supposedly allows for student input but which ultimately restricts the choices that might tell us the most about their writing ability. Reliability and validity do not give us ways to explain and legitimate the pedagogical practices we value in the writing classroom and our evaluations of the work our students produce. We either need a way to address the concerns for objectivity and consistency coming from educational measurement theorists and governing agencies, or we need to challenge them. Scholars working in portfolio assessment have not been able to do so successfully to-date.

## THE POWER OF OBJECTIVITY

An alternative theory for large-scale writing assessment must begin with the understanding that objectivity has been the primary driving force behind contemporary assessment. While other considerations, such as efficiency, have played a role, objectivity has been bolstered by the force of positivist science and consequently has carried the day. In fact, the story of writing assessment I have constructed here demonstrates how in large part writing assessment has taken particular forms specifically to avoid the subjectivity that positivist science has taken pains to erase. From Mann's praise of impartiality in 1845 to White's defense of the reliability of holistic scoring, the desire for objectivity has motivated the form of large-scale writing assessment in the United States.

During the last twenty years, however, postmodern theory has attacked objectivity, pointing out that objectivity begins with a stance that is as invested as any "subjective" position. These arguments, which I will take up in further detail in the latter half of this text, claim that in the final instance, there is no such thing as absolute objectivity. Consequently, most disciplines today, including the "hard" sciences, are questioning the nature of the reality their fields examine, and many are acknowledging that these realities are constructed through the frames of particular interests.

Some postmodern reflection exists in the assessment literature outside the field of composition. Egon G. Guba and Yvonna S. Lincoln's *Fourth Generation Evaluation* (1989), for example, argues explicitly for a "constructivist" or "fourth generation" evaluation that differs from existing and previous generations primarily in its focus on negotiation among all stakeholders in any assessment. These earlier generations, they argue, depend on positivist methodologies that cannot permit stakeholders to have a central place in evaluation, in large part because of the subjectivity they bring to any assessment practice. Constructivist evaluation finds both the purpose and the uses of evaluation to be constructed by all the stakeholders in the evaluation and thus breaks radically from positivist evaluation which measures "reality" independent of those involved in the assessment itself. Guba and Lincoln, however, are education scholars, and their work has been taken up only sparingly in composition studies.[13] Yet they offer an alternative vocabulary for writing assessment, which I examine in more detail in chapter three, and while it may not provide all the answers composition scholars are looking for—and Guba and Lincoln explicitly claim that it will not—it does provide suggested avenues for exploration.

The same is true of the work of Pamela A. Moss, whose work I return to in chapter four. In "Can There Be Validity Without Reliability?" (1994), for example, Moss argues that a hermeneutic approach to assessment could incorporate context in ways that reliability as a defining principle does not allow. Her work gets a bit more airplay than Guba's and Lincoln's, primarily from scholars like Huot and Williamson. But the limited exposure seems odd, considering that the constructivist model composition scholars advocate has been an explicit part of the discipline's theory since at least 1986 when Kenneth A. Bruffee's "Social Construction, Language, and the Authority of Knowledge" came out. Social constructionist researchers in composition, however, have focused primarily on the act of composing, and their theoretical work has not transferred well to assessment.

Kathleen Blake Yancey does some work toward bridging this gap. In her retrospective, "Looking Back as We Look Forward: Historicizing Writing Assessment," Yancey (1999) argues that writing assessment in composition studies has moved through a series of overlapping waves and that we are now in a space where compositionists have developed certain kinds of expertise that allow us in turn to develop better assessment practices. She claims that the pendulum of writing assessment theory has swung from a focus on reliability in the first wave to a focus on validity

in the third, and she points to a change in the idea(l) of reliability, away from statistics and toward interpretation and negotiation (1999, 491–2). Yancey's picture of the field tends to be hopeful, and she gestures toward a fourth wave in which any number of changes are possible. As her historical construction progresses, the terms "validity" and "reliability" appear less and less frequently and are replaced by concepts such as "interpretation" and "reflection" and "ethics." In this way, among others, her essay does important work, getting us to think about other approaches to assessment. However, Yancey does not directly challenge validity and reliability (although, admittedly, a retrospective may not be the appropriate forum for such a challenge). She points out compositionists' "reluctance at best, and aversion at worst, to writing assessment," and the way assessment is often "foiled against a teaching grounded in humanism" (1999, 495). Compositionists, she notes, do not want to be assessment people, even though assessment has always been bound up in writing.

The history of writing assessment as I have constructed it here suggests some reasons why composition scholars and other stakeholders might resist, either consciously or unconsciously, employing an alternative paradigm for assessment even as they embrace its models for other areas of teaching and scholarship. The testing of writing in the United States came into being during the same moment as the rise of positivism, the last quarter of the nineteenth century. Almost immediately, the pressures on writing assessment tended to push such evaluation toward increasing standards of reproducibility and scientific precision. Composition scales were developed explicitly in an attempt to regulate and standardize the scoring of student compositions—if not across the country, then at least within a district or school—and their mission was defined in these terms. But when composition scales failed in their self-defined mission, they opened the door for other methodologies. The result was objective testing, and between 1930 and the mid-1960s, there was no other game in town. During this time, the objectivist paradigm solidified, becoming the norm and dictating the terms of assessment.

The most recent methods—holistic scoring and portfolio assessment—look like significant breaks from positivist assessments, but they are not as long as they claim legitimacy through "validity" and "reliability." These terms, even in current usage, make sense only in an objectivist paradigm that not only acknowledges but also actively seeks objective reality. If change is integral to assessment, as it would be in a constructivist paradigm, for example, reliability becomes highly questionable and limited as a defining term. When holistic scoring and portfolio assessments use

these objectivist ideals to provide evidence of their worth, they remain solidly within that paradigm. The need for legitimation in the face of conflicting demands such as these tends to limit the exploration of alternative models for assessment and to keep especially large-scale assessment programs from straying too far—theoretically, at least.

The public nature of most composition programs further militates against embracing constructivist assessment or at least against relinquishing the objectivist paradigm. Universities tend to treat composition, like some other general education requirements, as common property, as subjects which are remedial in some sense and which can be taught by almost anyone on the staff. Such core requirements receive more scrutiny from those outside the university as well, and thus English departments and specifically writing teachers are blamed when graduates do not produce grammatically correct prose—the common understanding of the content of composition courses. "Hard" numbers demonstrating improvement or decline satisfy more people—including politicians and parents—than carefully articulated narrative explanations of a result or rationales for assessment methods. The objectivist paradigm provides writing teachers and administrators with a tried and true way to deliver these numbers under increasing pressure to demonstrate the success of particular programs.

Extricating assessment from objectivism is no simple job, nor is it entirely clear that compositionists should want to do so. Objectivist assessment grew out of a desire in part for more equitable and meaningful assessment, ideals few educators would reject today. Objectivity is thus both desirable and limiting. The problem for those considering alternative assessment paradigms is twofold: retaining these ideals while moving beyond the confines of objectivist principles. In many ways the latter is the more difficult task because, more than convenience or habit, objectivist principles are integral to assessment as it is currently conceived. Until an alternative theory of assessment becomes widely accepted in composition, anyone implementing, analyzing, or theorizing about a writing assessment program must speak to those values. Without an alternative vocabulary with which to talk about assessment, objectivist thinking will continue to direct the ways arguments about large-scale assessment are conducted.