Edward M. White

An Apologia for the Timed Impromptu Essay Test

Twenty years ago, the timed impromptu essay test was everywhere defended by English faculty as the most effective, responsible, and teacher-supportive assessment device available. Today, it is under attack from all sides as formulaic, unresponsive to the nature of writing, and destructive to the curriculum. For example, even as moderate and scholarly a team as the one from Miami University says about classes learning the writing processes, "a single-sitting, impromptu essay exam neither tests the skills these students have been taught nor supports the connections between learning and writing that a process approach assumes" (Black et al 9). Peter Elbow says essay tests give us "agreement about a faint, smudged, and distorted picture of the student's writing ability" (Belanoff and Dickson xiii). A recent NCTE Sense of the House Motion (No. 3, Nov. 1992) resolved "that the Council oppose the practice of claiming to measure a student's overall ability at writing by means of a single score on a single piece of writing produced at one sitting, and be it further resolved that the Council work to eliminate this practice." Even holistic scoring sessions, praised in the past as examples of collegial communities of teachers agreeing on standards, are now under regular attack at conferences by scholars as diverse as Peter Elbow and Richard Larson for supposedly forcing agreement to inappropriate standards by tyrannical methods.

Since, according to a recent survey (Murphy et al. 25), over 70% of colleges and universities that assess writing use some form of impromptu essay as part of their writing assessments—from placement of entering students to certification of departing seniors—these attacks must induce...
concern and guilt on the part of those participating in essay testing programs. For years, these faculty had been comfortable they were doing the right thing and now they find themselves wielding an instrument of the devil. Some are moving hesitantly into portfolio assessment, but for many others the new and substantial investment of time, energy, and funds required for portfolios are not available; the essay test itself has stretched resources almost to the breaking point. But is it worth the effort to maintain the essay test, with its limitations and faults, when cheap and easy multiple-choice tests are readily available?

The time has thus come for a sober reevaluation of the essay test, with attention to its virtues as well as its drawbacks, to where it may be useful as well as to where it is not. And the time has come for portfolio advocates, among whom I number myself, to recognize the important role essay testing has played in the past—and can still play—and to stop attacking essay testing as an unmitigated evil in order to promote portfolios, which can stand very well on their own.

The recent turn against essay testing is a natural result of conversion to a new and more convincing belief system which seems to demand denunciation as well as renunciation of the old one. Such conversions are not restricted to religion. We have seen a parallel pattern in the routine attacks on the old New Criticism, the close reading of the literary text, that open so many poststructural critical essays. Before the deconstruction or the social construction of the text can begin, some heretofore respectable critical reading from the 1940s must be pilloried as naive and wrong. The fact that our formalist forebears taught us to read with new care and attention is forgotten while their narrowness is treated with scorn. Similarly, the attacks on essay testing ignore its crucial value for the development of portfolios and its continued value in writing assessment, in an unseemly and often uncritical eagerness to promote portfolios as the new panacea for writing assessment.

At the outset, let me acknowledge that I was one of the principal proponents of essay testing during the 1970s, one of a small group of English faculty who (along with some personnel from testing firms) developed and promoted holistic scoring methods that allowed quick, economical, and reasonably reliable rankings of large numbers of test papers. From this distance, it seems clear that in our enthusiasm for essay testing (we too imagined that we had found the answer to all the problems in writing assessment) we were rather too careless on some issues and wholly neglectful of others; the recent attacks on essay testing have uncovered important problems, which I will not ignore in the discussion which follows. At the same time, it is also clear that the spread of large-scale essay testing brought about some major improvements in writing assessment,
along with some other, less clearly positive, changes in writing instruction. Portfolios are a natural outgrowth of essay testing and represent a clear advance in the assessment of writing under some circumstances; they also seem to be affecting the teaching and assessing of writing in positive ways. It is natural for portfolio advocates to stress its advantages over the essay testing that has now become familiar, even routine. But a wider perspective shows that the similarities between portfolios and essays are much more significant than their differences, since both call for *student production of texts* in response to assignments. Indeed, we might consider an essay test to be an abbreviated portfolio, or a portfolio to be a much expanded essay test. It also seems clear that one or the other of these methods of assessment may be most appropriate according to the particular needs of a particular situation and that either is preferable to multiple-choice testing in almost all cases.

The wide use of essay testing allows many of those now looking at writing assessment to ignore multiple-choice testing and its threat to writing. So we must remind ourselves that for most of our colleagues outside the English department and for almost all administrators, assessment *means* multiple-choice testing: evaluation of actual writing, whether on impromptu essay tests, term papers, or portfolios, is still generally seen as hopelessly subjective, unreliable, and arbitrary.

This wider perspective is particularly important, for many of the attacks on essay testing fail to take into account the way it developed as a response by writing teachers to the dominant multiple-choice testing in American education.¹ Even today, the supposed efficiency, objectivity, and economy of multiple-choice testing remain in the background of most discussions of writing assessment. In most situations, the choice is not, as much of the current debate would have you believe, between an impromptu essay and a writing portfolio, but rather between a multiple-choice test and some kind of essay test. The most recent survey of placement practices shows that “of all assessment methods being used for this purpose, approximately 48% are multiple-choice, 49% are timed writing samples, and 3% are portfolios” (Murphy et al. 22). Anyone with a role in writing assessment must keep in mind the multiple-choice specter that hovers just off stage; no stake has ever been driven through its heart to keep it in its coffin, however much it may be wounded.²

**Writing Assessment in Context**

Any assessment is, in one sense, a means of gathering information. That suggests we should be clear about what information we need and how we will use it before we decide about the means of gathering it. Another way
to conceptualize the issue is to think of an assessment as providing answers to questions; the questions need to be well formulated before we seek the answers. These ways of thinking about any assessment seem obvious enough, but very few assessment programs of any kind actually follow them. In most cases, the method of testing is the first issue decided instead of the last; the answer is sought before the question has become clear. Typically, some influential campus figure or group presses for a particular test (for example, a new measure of general education performance published and heavily promoted by some testing firm) or a fashionable new type of assessment (such as portfolios) as a valuable device in itself. But no assessment exists outside of its context, its uses, and its effects; no tests or assessment systems have value in themselves. Further, it is wasteful and intrusive to gather more information than we can well use and it is pernicious to use instructional funds to gather information that has no instructional purpose. And it is dishonest to use a test designed for one purpose (college entrance screening, for example) as if it provided information for another (placement into freshman composition). Before we decide on or accede to a college decision on any assessment method, we must insist that it respond directly to clearly defined and clearly stated goals.

The typical college placement test, for example, seeks information that will help students enroll in courses for which they are ready, so they have a reasonable chance to succeed. In English composition, that usually means a rough division between those who are likely to do passing work in freshman writing courses and those who need additional instruction before they have much chance of doing well. What kind of information is actually needed to make a placement decision?

Not much, in fact. Although the decisions about this information must be made on campus, with participation by the teachers who know what is in fact being required and taught, most American campuses find students ready for freshman composition if they can read pretty well and write complete sentences. More selective schools might look for more complex matters, such as an ability to write coherent paragraphs or to analyze complex texts; some open enrollment schools may be more concerned with vocabulary, fluency, and mechanics. But remember, we are speaking here of what is needed at entrance to freshman composition, which usually devotes a good bit of attention to paragraph and essay construction and more complicated reading. Most of us would be happy if our entering freshmen really knew how to predicate and read.

If this—or something like this—is all that the testing program is trying to discover, an essay test is perfectly appropriate. A test development committee can come up with test questions that will allow students to
demonstrate these abilities, with whatever variations are appropriate for the local situation, and 45 minutes allow plenty of time.

We can, in fact, learn a great deal from an impromptu essay and that is one reason that numbers of portfolio programs now include an impromptu essay as part of the portfolio itself. Among other matters, we can be sure that the student sitting and writing is the author of the work to be evaluated. But beyond this, we can take advantage of the focus and concentration that impromptu essays require. For instance, several of the portfolio programs described in Belanoff and Dickson use impromptu essays for self-assessment and reflection about the portfolio itself.

On the other hand, numbers of institutions are using short essay tests to examine students for entry to junior level courses or even as a condition for graduation. When we look at what those tests are attempting to discover, the inappropriateness of a brief impromptu essay (by itself) for that purpose becomes obvious. Among other matters, advanced or graduating students should be able to use sources intelligently to support—not substitute for—their own ideas, discover and revise complex arguments, show some depth of understanding of a topic, and understand the discourse community of a particular field. It is unlikely that a short impromptu essay test can provide us with much useful information about such matters.

My point is really simple: No assessment device is good or bad in itself but only in context. Only when we know what we are seeking to discover can we claim that a particular kind of assessment is appropriate or not. In the many situations that are appropriate for essay tests, we should proceed to use them without guilt—and without exaggerating what they are able to measure.

The Essay Test as a Form of Writing

I have already suggested the most important argument for an essay test: it is not a multiple-choice test. The difference between the worlds of the essay test and the multiple-choice test is vast, much greater than that between an essay test and a portfolio. The universe of the multiple-choice test is one in which all questions have a single right answer, embedded in a series of wrong answers. Where does such an artificial world exist, outside of the testing room? Furthermore, the right answer only is “right” because the creators of the test have declared it to be so by a circular set of rules. Those who have learned to succeed on such tests do so not by asking which answer is correct in the world or under various circumstances, but by choosing the one the test makers are likely to have chosen to fit the needs of the test. The multiple-choice test thus examines—along with its “content”—the degree to which the student can adapt reality to the needs of
authority. This indeed may be the reason that many such scores correlate well with success in college. The required submission to the world of that kind of test may also suggest reasons why minority groups score less well on these so-called "objective" tests than they do on writing tests (White and Thomas, Koenig and Mitchell).

But the only way to answer an essay test is to write—an obvious fact not much mentioned these days. Writing well on an essay test means recalling information, selecting an appropriate vocabulary, constructing sentences and paragraphs, and, somehow, having something to say. Every essay test shares the artificiality of all tests, but it does require an active response rather than the passive submission called for by multiple-choice examinations. While it is naive to imagine an essay test as a valid measure of all writing, it is disingenuous to ignore the fact that the production of writing for an essay is a wholly different activity than filling in the bubbles on an answer sheet.

Thus the second great advantage of an impromptu essay test is related to the first; since it is not a multiple-choice test, it calls for writing. As one who has led perhaps 200 essay test readings over the last two decades, I would go further: It often elicits excellent writing, memorable writing. How, you ask, can students turn out good writing with little time to reflect and no time to revise? Perhaps another question is the best answer to that one: How much reflection and revision goes into most student writing? I remember one cynic at a scoring session replying to a new essay test reader who was wondering at the high quality of some high school essays written in 45 minutes: "They’ve taken more time to write the test than they’ll give to their freshman themes." Even as we deplore the lack of revision opportunities for essay test writers, we should note that intense focus on writing for an hour or so can lead to valuable work.

In our passion for the writing process and in our quest for revision, we should not forget that some students can turn out fine work in an intense but short period of time. We need to be careful about defining "the writing process" as if there were only one or two right ways to proceed, and we need to remember that revision sometimes makes writing worse—particularly personal experience writing, whose power and voice tend to disappear as students fix up their drafts to meet what they take to be teacher expectations.

I am not arguing here that impromptu writing is better or more valid under all circumstances for testing than revised writing. Revision is important and many assessments will need to look at student ability to revise. And I do not mean to exaggerate the number of excellent essays that appear on large-scale essay tests; as with most large collections of student writing, including the creative writing contests many of us wind up judg-
ing, most of the work is dreary and formulaic. I concede that one serious drawback for essay testing is that it does define writing as first-draft writing and I have elsewhere argued for portfolio assessment at the college graduation level because it allows for a much fuller definition of writing (White, "Assessing Higher-Order Thinking"). But I am disputing the assumption that the impromptu essay is in itself an invalid measure or that it is to be scorned because it does not test revision. If we accept the need for assessment and if we recognize the need to compromise the perfect assessment in the light of human and financial cost, we can claim that first-draft writing is a form of writing and one that can give useful information about student ability. Furthermore, it does allow students to focus intensely on writing for a limited period of time and, if the question draws out the creativity of at least some students, can elicit some very good writing indeed.

Although an essay test can reflect current writing and reading theory, lead to genuine writing, and yield needed information, I do not mean to exaggerate the virtues of impromptu essays. Its problems have become clear. The impromptu essay test—like all tests—restricts the world of the student, who must write under time pressure to someone else's topic and scoring criteria. The audience is almost always that looming unknown, an English teacher who will be grading the test—even if the directions specify a particular audience (a letter to a relative or to an editor, a report to the boss, etc.). A test-wise student knows that, while Aunt Sally may not care a whit about spelling or punctuation, no English teacher will appreciate a mechanical mess. When test development is weak, the essay question may be just as artificial and just as distant from student concerns as any multiple-choice test item; and when test preparation is divorced from serious writing instruction such mechanical response patterns as the five-paragraph theme may make the response almost as passive. Careful test development can provide clear and stimulating topics. But no matter how excellent the question, it is only one question (on rare occasions two) and therefore restricts the multifold domain of writing to narrow categories. The standardized test conditions strip natural context from the writing, allow no collaboration or preparation, and disallow more than token revision. While I maintain that the essay test remains far superior to multiple-choice testing, we must acknowledge that impromptu essays leave much to be desired and (again like all tests) yield limited information. While closer to the definitions of writing used by teachers than a multiple-choice test, essay tests simplify those definitions in ways that many teachers find oppressive. Clearly, essay testing is a compromise and its limits must be seen clearly.
The Message of the Essay Test

Every assessment defines its subject and establishes values. Since assessments have a powerful effect on what is taught, we are properly concerned that the values and definitions embodied by assessments support teaching, whatever else they may accomplish. In this regard, the essay test message is somewhere between that of the multiple-choice test and that of the portfolio. One of the major objections to the use of multiple-choice writing tests is that they devalue writing; the way to prepare for the writing test is to develop techniques (such as probabilities of guessing) of filling in the correct bubble. One of the major advantages of the writing portfolio is that it values a variety of writing experiences; the way to prepare for the writing assessment is to compile examples of good writing. The effect of the multiple-choice test is to diminish the importance of writing; the effect of the writing portfolio is to increase the amount and scope of student writing. Where does the essay test fit between these two extremes?

I have no doubt that the essay test supports the teaching of writing, even though it measures impromptu writing products and hence is less effective than most portfolios. The most obvious example is the writing done in Advanced Placement classes in high schools all over the country. While too many AP teachers restrict the writing in these classes to practice for the essay tests their students must take, not all do; meanwhile, the students are actually writing, on demanding topics, to clear criteria using high standards. Some college teachers disapprove of AP because of its relatively restricted view of both reading and writing. Mahala and Vivion have recently argued that the present exam is not valid, most particularly in its multiple-choice portions, which do seem to be out of touch with current views of literature and rhetoric. The authors also claim, less convincingly, that the scores from the AP exams should not be accepted because its assumptions about writing in the essay portions "are at odds with our writing program's curriculum." That curriculum assumes "that writing processes that make the search for meaning possible are not necessarily available on demand and are not always even available to conscious control" (Mahala and Vivion 49). But that argument makes the common mistake of confusing a curriculum with an assessment; the exam is seeking information about students who do have conscious control of their writing process and can show it on an essay test. Only such students deserve credit by examination; the rest will appropriately enough go through the curriculum. As with many exams (and, indeed, many portfolios containing only products) the assumption is that a high-scoring product will actually reflect the refined process that led to the product; one measures the process by
sensitively evaluating a product. Thus the best and only useful preparation for AP essay exams is to write and to develop a writing process that allows good writing to appear on the test.\textsuperscript{3} When we look at the arguments that have led to the appearance of essay questions on other large-scale national tests, such as the General Education Development (GED) high school equivalency test or the Medical College Admissions Test, the message function of the essay test figures prominently: the medical schools, for instance, use the essay to urge undergraduate premeds to broaden course selection beyond the sciences.

Portfolio advocates have a convincing argument that the natural and broad preparation necessary for their assessment is by all odds the best message; students who must present portfolios seek out writing experiences. But I wish we could make this case without demeaning the value of the essay test message, which has so powerfully improved the teaching of many subjects. Instead of focusing on the bad examples of some narrow and illiberal teachers drilling students in five-paragraph themes (we must wonder how such teachers will debase portfolios in time), we might better praise the minority of overworked and underpaid school teachers who are teaching writing at all under disheartening conditions. Many writing across the curriculum programs in the schools and colleges labor to convince teachers to include even a little writing in their courses and the replacement of a multiple-choice test with an essay test can be a major breakthrough. There is no debate about portfolios being superior to essay tests in principle; multiple measures are always better than single measures. Perhaps portfolio advocates could help those responsible for essay tests to expand the scope of those tests to include some forms of the wider sampling that distinguish portfolios.

The Timed Impromptu Essay Test Vs. Portfolio Assessment

Portfolio assessment is valuable because it potentially reflects the wide range of writing that we teach and the writing processes as well as the writing products of the student. Further, a student compiling a portfolio takes on responsibility for his or her own work and is usually asked, as part of the portfolio, to perform some kind of self-assessment. The portfolio assessment uses multiple rather than single measures written over a prolonged period of time and under classroom rather than testing circumstances. Furthermore, a portfolio program has much larger instructional effect (usually for the good) than less intrusive assessment procedures, since it brings outside assessment directly into each teacher's classroom; no other assessment procedure looks so pointedly over the teacher's shoulder
at particular assignments, student responses to those assignments, and
teacher responses to what their students have written. For all of these
reasons, and for many more, a portfolio assessment is potentially a more
valid and more meaningful measure than a single impromptu test, with
more positive uses for instruction and less negative conditions for the
student.

These are persuasive arguments and have led to the present wave of
interest in portfolio assessment. Yet we should not minimize the problems
of implementing a portfolio assessment. Portfolios combine a collection
activity, with all the difficulties of dependable and meaningful collection,
and an assessment activity, with profound difficulties for consistent and
meaningful measurement. Under the right circumstances, these difficulties
can be met and the assessment will have great value for the students and
institutions involved. (See Roemer, Schultz, and Durst, who describe the
systematic and careful replacement of a less valid essay exit test by a more
valid and useful portfolio assessment). But, as with any complex and high
quality operation, the financial and human cost is high enough to make
reasonable people ask if every assessment should be a portfolio assessment.
Assessment should support but not replace instruction and we need to use
resources wisely. As I argued earlier, the context and purpose of the
assessment, not abstractions, should determine the assessment device.
Thus we need to look with some skepticism at some of the claims made by
portfolio advocates to see which claims hold and to determine if they are
worth the price.

Those newly involved in portfolio assessment frequently marvel at the
positive force it brings to faculty debate. They speak of energetic discussion
of the goals of composition courses—or even of the entire college curricu-
lum. They note that a portfolio reading is a collegial enterprise with its own
energy, focusing upon teaching and learning, often professionally reward-
ing despite long hours and sloughs of boredom. They also note that the
effects of such an enterprise last long after the last score has been given,
altering teaching patterns and ways of responding to writing.

All of this is true. But it is equally true for locally-developed essay
testing, as two decades of experience have shown. We should not accept
the assumption that essay tests are necessarily invalid and impersonal, run
by outside forces; any assessment can be run well or badly, locally or by a
commercial testing firm. Whenever a local assessment—whatever the
kind—becomes the responsibility of faculty, the usual small talk and pri-
ivate edges begin to disappear before the large professional task. The re-
sponsibility of deciding what we shall measure and how is profound, and
even cantankerous colleagues sometimes find themselves listening to each
other, wavering on long-held views, agreeing to agree. Whether the assessment requires essays or portfolios, the benefits to faculty who participate are long-lasting and powerful.

Reliability and Validity

Validity means honesty; the assessment is demonstrably measuring what it claims to measure. Reliability means consistency; the scores are demonstrably fair. It is possible to have reliability without validity, by consistently scoring a test which fails to measure up to its claims. That is the usual complaint about many multiple-choice tests, whose scoring by computers is absolutely consistent, but whose results may relate more highly to parental income and the dialect spoken in the home than to student writing ability. Though we can have reliability without validity, we cannot have validity without reliability, since without reliability our results are essentially meaningless and meaningless scores cannot be validated. Statisticians insist that reliability is the upper limit for validity, that no measure can be more valid than it is reliable. For this reason, we need to discuss reliability before we dare claim validity.

Essay test reliability has been a major focus of assessment research for the past two decades. Beginning with Paul Diederich's discovery that even selected readers without scoring criteria or controls will give all essays all possible scores, procedures to increase reliability have received widespread attention. The development of holistic scoring—with its paraphernalia of scoring guides, sample papers, chief readers, table leaders, and complex computer record-keeping—has led to the present situation: A well-trained and experienced cadre of readers can score many thousands of student papers on a six-point scale with about 95% agreement on scores within one point. This is not the perfect scoring consistency of the computer, but close enough for useful measurement. (See Williamson and Huot for the most recent studies.)

This scoring reliability comes with both a financial and professional price. Readers, table leaders, chief readers, aides, computer technicians, and facilities for a controlled reading require funds. The usual essay scoring session costs roughly $10 a paper, depending on student level, length of papers, pay of readers, and so on. For some programs, the financial cost is easier to handle than the theoretical price readers must pay: Essay raters must submit their independent judgments to group standards, the final arbiter. The forced agreement, for the sake of fairness to the students, grates on some faculty, who will complain that such agreement does not reflect the real world of disagreement on standards. Other faculty, with whom I am more in sympathy, argue that forcing faculty to test their
grading standards by comparing them with those of their colleagues is a reasonable and healthy exercise. Besides, agreeing on standards for a particular essay question and student group, for a specific assessment purpose, is not the same as agreeing on more global concepts of writing, which no experienced academic would ever imagine university faculty to be able to achieve. Unless agreement on standards, expressed theoretically in a scoring guide and practically on sample papers, can be reached, scoring reliability will be too low for the scores to be meaningful. And unreliable scores by definition mean an invalid assessment.

However, reliable scores are harder to achieve than many essay test proponents realize. Even though we know how to obtain consistent scores for essay tests, scoring reliability is not the same as test reliability. The pilot project of the Medical College Admissions Test (MCAT) showed that even the extremely high reliability of a single expert scoring does not produce reliable test results. When some 600 students repeated the spring MCAT essay test in the fall, with each test scored consistently, the second score for many students varied considerably from the first—that is, reliability was relatively low (about .5). Even though each score was a dependable measure of the writing produced on the particular writing topic, there was no way to tell which measure accurately described the underlying student ability reflected by the writing. The problem lay in putting too much weight on a single measure. An essay test is essentially a one-question test. If a student cannot answer one question on a test with many questions, that is no problem; but if he or she is stumped (or delighted) by the only question available, scores will plummet (or soar).

The MCAT solution was to replace a single longer essay test with two 30 minute essays. The test score derived from combining the two essay ratings was much more reliable when compared with a second essay test score (now about .75; see Mitchell and Anderson). This result confirms many other studies, all showing that confidence in a single test rating is usually misplaced and overstated. The essence of reliability findings from the last two decades of research is that no single essay test will yield highly reliable results, no matter how careful the testing and scoring apparatus. If the essay test is to be used for important or irreversible decisions about students, a minimum of two separate writing samples must be obtained—or some other kind of measurement must be combined with the single writing score.

This problem with test reliability for essays, no matter how accurate scoring reliability may be, poses some hard decisions for those administering essay tests. If the test is being used for a barrier—to freshman composition course credit, upper division advancement, graduation, etc.—the high stakes disallow the use of a single essay to obtain a writing score. A
lower stakes test might use a single sample, as long as everyone acknowledges its relative unreliability. For example, one might continue to use a single essay score for freshman placement as long as course instructors can move students misplaced by the test during the first week of classes. While it would be better to have a more reliable test, it may not be worth the cost, particularly if the funds must come from instruction.

The painfully won and documented scoring reliability of essay tests is now in the range of credibility to influence many decisions—even so tense and litigious a decision as admission to medical school. While that reliability is not perfect, it is high enough to allow essay tests to claim validity as a direct measure of writing ability. This validity is much more limited than many of us would like and not nearly as encompassing as many of us claimed two decades ago. But without reliability, claims for validity—such as are being made for portfolios—ring hollow and false. Portfolio assessment may eventually be able to demonstrate reliability, but studies have only recently begun on ways to achieve it. The difficulties are enormous, probably many times those of single essay scoring, but some programs (such as the ones at Miami University of Ohio, the University of Michigan, and New Mexico State University) are coming up with encouraging statistics. However, an evaluation of the scoring of the Vermont Portfolio Assessment Program, the most ambitious use of portfolios in the schools, is not promising: "The overall pattern was one of low reliability, however, and in no instance was the scoring highly reliable" (Koretz et al). Portfolios contain multiple samples of writing, which should increase test reliability, even as it makes scoring reliability more difficult. We should not follow the lead of those who, despairing of achieving reliable scoring of portfolios, attempt to discredit the concept of reliability—as if we could only achieve honest measurement by abandoning fair scoring. The future of portfolios depends on the development of reliable scoring procedures, which will almost surely be adapted from those now known to work for single essays.

Essay test validity, once its vaunted strong point, has been under particular attack in recent years. In the 1970s, we spoke confidently of the writing done by students on an essay test as "real writing"—which, in comparison to multiple-choice testing, it surely was. However the advent of portfolio assessment has shown how narrow the essay test definition of writing is. We need now to say that the impromptu first draft writing of an essay test represents one kind of writing which may or may not reflect a student's ability to produce other kinds. The score on an essay test may reliably show how well the student wrote on the particular question in the time available under the conditions given on the day of the test. Is such a limited definition "real writing"? Well, it is certainly not "unreal writing"
and it is certainly writing, but, like all writing and all tests, it is artificial
and subject to many influences and interpretations.

Those without much experience with assessment may find all of these
cautions about reliability and validity to be a severe condemnation of the
direct measurement of writing, particularly essay testing. But the fact is
that all measurement of complex ability is approximate, open to question,
and difficult to accomplish. For instance, a little-known statistic hidden in
the back of obscure test materials, the Standard Error of Measurement,
often states stunning possibilities of inconsistency despite the accuracy of
computer scoring. Naivete about the inevitable approximations of all
assessment instruments causes almost as much difficulty for assessment as
the opposite mistake, distrust of all measurement. To state that both
reliability and validity are problems for portfolios and essay tests is to state
an inevitable fact about all assessment. We should do the best we can in
an imperfect world.

Cost

We need not say much more about cost, but we should emphasize
the complexity of cost calculations. Multiple-choice tests are cheap to
score, but they yield little of value in writing assessment. Portfolios are very
expensive to collect and score, but their yield is great. Essays again lie
between the two. The most rapid portfolio scoring yet reported is the
one at Miami University, which is able to achieve about five scores an
hour from trained readers. The average essay test reader can score about
twenty-five essays an hour, reading 45-minute papers, after a relatively
short training period. That is, essay scoring costs about five times as much
as multiple-choice testing and portfolios cost about five times as much as
essays. Are they worth it? Surely the answer depends on the source of the
funds and the uses of the assessment.

When we deal with the issue of cost, we need to take pains to point out
that cost by itself is meaningless. What matters is cost effectiveness, what we
get for the money spent. Something that is cheap and useless is less
cost-effective than something that is expensive but highly useful. For
several decades, proponents of essay testing have argued, with consid-
erable success, that the development and scoring of such a test delivers
much more than a score: heightened student interest in writing, intense
faculty discussion of standards and goals, in-service training in question
development, curriculum revision, improved teaching and learning, and so
on. These valuable by-products of a careful essay testing program add
substantially to its cost effectiveness, making it a bargain despite its ex-
pense. Recently, proponents of portfolios have been making much the same arguments, asserting that the higher cost of portfolios brings even greater benefits: direct links between teaching and authentic assessment, richer definitions of writing, thorough revision of the curriculum, valuable documents for students for counseling and job application, student responsibility for learning, and the like. This is not the place for a discussion of portfolio financing, but the widespread interest in portfolios despite their cost suggests that these arguments are also persuasive.

When an institution seeks to drop an essay test and replace it with a multiple-choice test, the usual argument stresses cost savings. It is up to the faculty to point out that a cheaper test might well have an exorbitant cost if all it produces is a list of scores, and relatively invalid scores at that. A careful assessment profoundly linked to teaching and learning can be shown to be a best buy. We might also delicately point out that inappropriate, unreliable, or invalid tests could become enormously expensive if challenged in court by students.

Finally, we must concede that essay testing, valuable though it is, has important limitations—in definition of writing, in message to students, in reliability and validity, in effect upon teaching. Given a choice, under most circumstances a portfolio assessment is clearly preferable, for all the reasons I have suggested here and many more. But institutions that cannot afford the money, time, effort, and uncertainty of a portfolio assessment should not retreat to multiple-choice testing, and we must be careful to allow the essay test option in such cases. Portfolio advocates should consider an essay test with one or two questions to be an abbreviated portfolio, one which might be expanded when the resources become available. When we point out the limitations of essay testing, we really have a professional obligation to note that all assessment instruments, including portfolios, are compromises and contain limitations. But the essay test is not a bad compromise, with all its shortcomings, and it still has much to offer a world in which most testing has become a negative, passive, and detrimental intrusion into the learning process.

Acknowledgments: I thank the following CCC consulting readers, whose thoughtful comments on early drafts of this article enriched its many revisions: Pat Belanoff, Russell K. Durst, and William Smith.

Notes

1. I describe this history in considerable detail in "Holistic Scoring: Past Triumphs, Future Challenges" (Williamson and Huot 79-106). I have also adapted this material for the final chapter in the new and expanded edition of my Teaching and Assessing Writing.

2. The most recent example of the replacement of writing by fill-in-the-bubble testing comes from an excellent private insti-
tion, DePauw University in Greencastle, Indiana. Its well-developed essay placement program, which had a valuable research component and substantial campus support, has now been abandoned to be replaced by multiple-choice tests. DePauw is also returning to an old idea, that placement of students is discriminatory and leads to negative grouping—that the attempt at a solution worsens or even creates the problem. This idea, which is recycled about every 30 years, has also been articulated recently by Peter Elbow in conferences and by Richard Haswell in *Gaining Ground in College Writing*. While there is a better argument for this “mainstreaming” in small or selective colleges than in large universities, I am concerned about the loss of opportunity it represents for poorly-prepared students. We have powerful evidence from statewide programs in California and New Jersey that carefully designed remedial/developmental programs help many students succeed who would otherwise fail.

3. I am not defending the present AP exam, which clearly needs revision along the lines suggested by Mahala and Vivion. But I do endorse the concept of credit by examination, as they do not, and I am convinced that AP has been a powerful force for improvement in the school curriculum, particularly in its emphasis on writing.

4. The Standard Error of Measurement of the verbal portion of the SAT is approximately 30 points.

**Works Cited**


