Marie Jean Lederman

WHY TEST?

Editor's Note: *This article appeared originally in an essay collection on writing assessment published in 1986. The full citation is given in the permission statement below.* JBW *is reprinting this article because we feel it provides useful background and texture for the dialogic essay written especially for* JBW *by Gao Jie and Marie Jean Lederman (see article immediately following this one). We feel also that on its own it offers an indispensible perspective on writing assessment in the United States.*

Why do we test? Some of us test because we believe we must. More of us test because boards of regents or trustees, state legislators, or high-ranking college administrators have mandated testing programs. In the mid-1980s in America, testing has become the flag raised by the troops of the Land of Academic Standards.

Today's strong belief in assessment ranges from the "quick fix" of tests in popular magazines to formal examinations in schools. The city of Minneapolis is a striking example. In its 1984 attempt to tighten academic standards, it was the first school system in the country to require competency tests for promotion out of kindergarten. To ensure preparation for testing at this level, the business community is busily developing computer materials such as Program Design's Baby's First Software.

America appears, at this juncture, to be a particularly test-happy culture. But what seems to be an especially American, especially contemporary phenomenon is far from unique to this one place and this one time. Today's spur to testing may be boards of regents or trustees, legislators, or local administrators, but the reasons we test and the inevitable problems involved in testing have roots that touch the beginnings of social activities.

To understand why we test today, it is instructive to go back to reasons why people throughout history and throughout the world have relied on tests. A look at other cultures and their tests provides a useful historical

---

From Writing Assessment: Issues and Strategies, *edited by Karen L. Greenberg, Harvey S. Wiener, and Richard A. Donovan. Copyright* © 1986 by Longman Inc. Reprinted by permission.

Journal of Basic Writing, Vol. 7, No. 1, 1988

DOI: 10.37514/JBW-J.1988.7.1.05

perspective on our own motivations for testing, our testing procedures, and the inevitable limitations of any tests we create.

Perhaps the earliest tests were *rites de passage*, tests that inducted adolescents into adulthood. These rites not only marked a sexual coming of age but also marked admission into the culture, values, and mores of the group. According to Otto Rank, they were educational experiences that reconciled, for boys, both sexuality and education by deferring the boy's formal education to the time of puberty. The boy's initiation marked the passing of the role of education from a person (mother) to the community; *"in place of a human being as a pattern of education,* a collective *ideology* appears as *the education ideal"* (246). Basically, such tests permitted movement of both girls and boys from one stage to another and were inherent in the education of all members of the group. Of course, the nature of the tests varied, depending on the values of the group. These *rites de passage,* marking a transition from one stage to another according to specific tasks performed, might be seen as harbingers of proficiency tests like the "rising junior" examinations given by some colleges today. These "rising junior" examinations seek to establish a set of tasks beyond course grades that are "external" verification of students' abilities to meet the standards of the group they wish to join.

If attaining membership in a group was one early function of testing, another was the attempt to sort people or to choose the best people to perform specific tasks valued by a group. The Chinese invented the examination, "one of the more controversial of their contributions to the world, which many centuries later adopted this method of determining qualifications" (Heren et al. 121). In China, the written examination system began in the Sui dynasty (589-618). The Chinese attempted to create a system of competitive examinations for government positions, precursors to our modern civil service examinations.

By A.D. 1370 these examinations had striking similarities to writing assessment examinations today:

> *Every three years competitors successful in the district examinations assembled in the provincial capitals for three sessions of three days and three nights each. Compositions in prose and verse revealed the extent of reading and depth of scholarship. At this level, penmanship did not count, since a bureau of examination copyists (established in 1015 A.D.) reproduced the papers in another hand before they were evaluated by two independent readers, with a third reader to receive and reconcile the sealed grades. (DuBois 4)*

In attempting to rank candidates on the basis of demonstrated merit, the examiners in China faced many of the problems that we face in designing similar assessment tasks today. One problem in essay testing now is the question of the influence of handwriting in judgments that readers make about the quality of an essay. This question seems to have been solved, at least to the satisfaction of the Chinese examiners. By rewriting candidates' papers, they ensured that handwriting would not "count" (DuBois 4). An alternative explanation, however, may be that

the decision to copy the papers was made to conceal the identity of the examinees. Other historians note that in addition to using numbers instead of names on the examination papers, papers were copied to ensure that the examinees' identity would remain unknown and therefore would not influence the readers (Fairbank, Reischauer, and Craig 189). Today's examiners, similarly, seek to maintain the anonymity of examinees through substitutions such as social security numbers or other codes on student papers.

An even more striking parallel with today's concerns about instruments for writing assessment was the early recognition of the problem of tests establishing fixed forms and of the relationship between those fixed forms and the creativity of the examinees. By 1487 in China a specific form for writing examination papers was adopted, "under eight main headings, with not over 700 characters in all and with much use of balance and antithesis. This was the famous 'eight-legged essay' style, later denounced as imposing a tyranny of literary structures over thought" (Fairbank et al. 190). Some scholars now see this examination system as having "degraded education and made it a mere appendage to the examination system" (China Handbook 4). Today we continue to worry about whether or not the format of an essay examination will have a negative effect on students' creativity and thinking or, worse, that our tests may become more important than our curriculum.

Another question we debate is frequency of retesting. How often should students be asked to repeat tests that they have not passed? According to Scharfstein, the answer in nineteenth-century China was so many times that "many candidates sat for these examinations for twenty or thirty years or more. At the age of eighty or ninety, candidates who had failed repeatedly might be given a consolation degree. They were failures, but honorable ones" (17). Few of today's colleges exhibit either such patience or such compassion. Neither, for that matter, does the rest of our culture.

An additional problem is the control of cheating. As one expects when the stakes are high enough, there may be desperation on the part of some of the candidates. In nineteenth-century China, for example, "expert stand-ins were hired" or "clothing was lined with thousands of microscopically written essays to which the 'padded' candidate had an index" (Scharfstein 18). Soldiers inspected the candidates for hidden papers, sometimes going "so far as to cut open dumplings in order to examine their bean-jam fillings" (Miyazaki 44). Despite these attempts, in certain periods, cheating was rampant.

Perhaps the most fundamental question troubling testmakers throughout time has been the question of equity. After all, the assumption of the civil service tests in China was an assumption of the basic good of a merit system. Whether tests are designed to mark a transition, to assess specific knowledge, or to sort candidates, the question of equality of chance to pass the test is universally present. The attempt that the Chinese made, over 1300 years ago, to sort candidates according to merit was admirable in theory. The reality, however, differed, for despite the attempts to make each examinee equal to all others, the system still

favored the sons of the rich. These examinees went to national schools at the capital. Moreover, many of these students could afford tutors and came from "scholar-official" families, which afforded them the additional advantage of a role model at home (Fairbank et al. 104, 190). Thus in the Chinese merit system, social class and wealth made some examinees more equal than others. Needless to say, the problem of equity in theory and reality persists in a variety of forms today.

We find ourselves kin to the examiners in China thousands of years ago, and as we move through the history of educational testing, we see other similarities in the examinations for university degrees awarded to the candidates of the first Western universities. Here the earliest examinations were oral; written examinations began in the thirteenth century, several centuries after the introduction of paper to the West. As Fairbank notes in *Chinabound*, "Europeans ... had argued in their universities for hundreds of years before Gutenberg while Chinese scholars had been using paper, brush, and printed books all the time" (372).

Still later the Jesuit order, founded in 1540 by St. Ignatius of Loyola, pioneered in the systematic use of tests in education. They used written tests both for placement of students and for ascertaining proficiency after instruction. In 1599 they published their statement of procedures for examinations in the lower schools. While some of the procedures seem quaint, others have a decidedly familiar ring:

> *The writing should be done in a style befitting the grade of each class, clearly, and in the words of the assigned theme and according to the fashion prescribed. Ambiguous expressions are to be given the less favorable meaning. Words omitted or changed carelessly for the sake of avoiding a difficulty are to be counted as errors.*
>
> *After the composition is finished, each one, without leaving his place, should diligently look over what he has written, correct and improve it as much as he may wish. For, as soon as the composition is given to the prefect, if anything then has to be corrected, it should by no means be returned. (DuBois 9)*

The strictures to be specific, to avoid ambiguity, and to proofread the paper have a timeless quality and are reminiscent of directions given to students for many large-scale essay examinations today.

By the middle of the nineteenth century, both oral and written examinations were routine in England, on the Continent, and in the United States, and written examinations were recognized "as an appropriate basis for important decisions: who should be awarded degrees; who should be permitted to exercise a profession, such as law or teaching or medicine; and who should serve in a government post" (DuBois 10).

In the nineteenth century in England, various refinements of the grading procedures for essay examinations were developed. DuBois notes that in 1864 the Reverend George Fisher of Greenwich, England, collected samples of academic writing and arranged them in a "'Scale Book' with assigned values from 1, the best, to 5, the poorest. Intermediate

values were indicated by fractions. Work by any student could then be graded by direct comparison with a set of specimens arranged in order of merit, thus providing a fixed standard of grading in each of the subject matter areas" (69).

Slowly, procedures were developed for measuring what students had learned by examining their writing. Fisher's "scale book" made explicit what was implicit in the minds of the examiners. Similarly, many educators who direct writing assessment programs today believe that it is important to illustrate raters' criteria through "scale books" that illustrate each point on the scale with real examples of student writing.

As we look at the growth of testing, we note that throughout history "whole" tasks were the rule: tasks performed as part of initiation rites and lengthy oral and written responses to questions. It is only in recent times that we have developed the notion of indirect measurement. When multiple-choice tests—easier to score and administer—arrived, we greeted them joyfully:

> *A great stimulus for the growth of educational measurement was the invention of the multiple-choice item, first used extensively in the Army Alpha. Educational test makers soon discovered that an item consisting of a clearly written stem, followed by four or five alternative answers, of which one is correct, provides a flexible format for the measurement of both knowledge and skill. (DuBois 73)*

The 1920s saw an explosion of such test construction for use in the schools and colleges. Not surprisingly, "Instructors liked the 'new examinations' because they were far more comprehensive than earlier methods of testing and because the chance of personal favoritism influencing scores was practically eliminated" (DuBois 76-77).

In 1900 the College Entrance Examination Board was founded to provide the country with a systematic testing program. Traditionally, only essay examinations had been used for college admissions, but after the development of the multiple-choice format during World War I and the uses of objective testing at Columbia College, objective tests were introduced into the board's testing program (DuBois 125). Varieties of other testing programs, such as the National Teachers Examination, soon began. In 1947 the three major education groups involved in testing, the American Council on Education, the Carnegie Foundation for the Advancement of Teaching, and the College Entrance Examination Board, founded the Educational Testing Service (Ebel 22). Multiple-choice testing was in.

The multiple-choice test has become so firmly entrenched in American life that it now seems revolutionary to call for "whole" tasks such as writing samples. But we must remind outselves that our immediate past—a mere half century—is hardly the whole of human history. Short-answer tests, which permeate popular culture in our magazines, are but one example of a pervasive societal quest to find simple, quick answers to complex questions.

There are many other examples. Television has woven the short-question, short-answer format tightly through our lives, not merely through quiz shows and sitcoms but through news reporting itself. Nightly, much of life is also reduced to "And what did you feel when you saw the body?" "I felt scared." Sixty-second spot commercials first became 30 and are now 15 seconds long. Worse, in classroom after classroom, educational level after educational level, short questions and short answers have become the norm. As John I. Goodlad asserts, students spend most of their time listening, some of their time reading short passages and writing short responses to questions on quizzes, and virtually none of their time reading or writing anything of some length. The destructive nature of the short-question, short-answer mode of living is apparent: not all of life's complexities can be summed up in one-sentence questions, much less one-sentence answers.

Despite the advantage of short-answer tests—the skills and knowledge that can be sampled and the ease of administration—a fundamental criticism remains. What many people consider to be the most important goal of education, coherent thought and expression of that thought, simply cannot be measured by multiple-choice or short-answer tests. Clear thinking and clear writing are inextricable. Writing makes us accountable in a way in which neither the spoken word nor short-answer tests do.

If we were to agree that coherent writing, which both produces and reflects thoughtful understanding and analysis, is the primary goal of education, the question of how to assess it would be easier to answer. But obviously we are not, as a group, in agreement on the primacy of writing in education, for both anecdotal reports and surveys tell us of the increase in both multiple-choice and short-answer testing in courses throughout colleges and universities. Even though most college faculty members know that they get a different kind of information about students' knowledge and abilities from essay tests than from short-answer tests, short-answer tests continue to proliferate.

A recent interesting experiment conducted with undergraduates at Florida International University supports the value of learning by writing. Students were divided into groups and were given a 4800-word passage to read. Each group was told to expect a different kind of test: an essay, multiple-choice, "memory," or some other unspecified kind of test. All the students took the same test, which included both multiple-choice and short-answer items. Students who were told to expect an essay test did better even on the multiple-choice items. The researchers theorize that when students prepare for an essay, they "take a broader focus" and try to organize facts by integrating them into a larger context. This kind of preparation apparently aids recall of the specific details needed to answer the multiple-choice questions (Cramer 17). Although research is not conclusive, it is hard to believe that teachers have not acknowledged the results of this study simply by intuition, if only from memories of the way in which they, as students, prepared for essay tests.

This point brings us back to the original question, Why test? The question must be answered—and with more than a short answer—before we

can discuss assessment instruments. Most English teachers would immediately say that we test to place students, to diagnose specific strengths and weaknesses so that we can help writers improve, to determine growth, and, finally, to assess either competency or proficiency. Some would say that we test so that we may design courses that will help students to become better writers. A few would add that sometimes we test students to determine whether our courses have succeeded or failed.

But the more fundamental question is, What, as a society, do we value? Is the ability to write a critical skill for success in our culture? If so, assessing student writing is an appropriate ritual. What form should that ritual take? Our ultimate goal should be to improve teaching and learning. Yet testing, which should be an outgrowth of and subordinate to curriculum, in reality often drives curriculum. Therefore, our choice of assessment instruments is crucial. If we do not want to encourage students in writing classes throughout the country to sit in classes and fill in blanks in workbooks or on computer screens, we will not use short-answer or multiple-choice tests. If we want to signal to faculty in both secondary and postsecondary institutions that the business of a writing class is writing, our assessment instruments will be essay tests.

Faculty members in departments other than English bemoan the fact that students cannot write. When pressed for an explanation, teachers say that students do not know how to isolate and stick to an idea, develop that idea, and illustrate it with specific examples. They talk less about surface and mechanical errors (the elements that are measured by short-answer tests) than about issues of logic, coherency, and detail. Short-answer tests are not our answer if what we want is a primary educational focus on thinking skills rather than editing skills.

A clear relationship exists between the curriculum we teach and our assessment instruments but we should not assume a total overlap between teaching and testing. No test, whether in a political science, biology, or writing class, can tap the entire domain of what the student has learned during an entire semester's work. No single instrument can deliver that kind of information.

A current example of the simplistic assumption of the complete overlap between curriculum and testing is the popular cry, "We teach process, but we test product." Like the 15-second spot advertisement on television, the complaint has a catchy ring but masks the complexities of assessment. Of course, the best teachers do help students learn something about their own writing processes, to overcome the points in their writing processes at which they are hopelessly stuck, to expand the repertoire of skills that students use when they write, and to learn the patience needed for creation and the joy of tinkering with their own prose. But in the end, it is a lie to tell students that "product" does not matter. As readers, for example, you are not interested in the 20-odd drafts that resulted in this chapter. The brilliant insight that may have flourished briefly before fading in the course of the writing process is of no use to anyone except, perhaps, the writer. What is altered does not matter to the reader, nor does the ease with which the writer composes. In the real world, product is all we can share with each other.

In an idealized universe, there is unending time for vision and revision. Nevertheless, curricula in our writing courses should allow time for students to explore many types of writing, from the quick and largely impromptu prose that most writing tests demand to the longer, more reflective essays for which students will have days or weeks to imagine, plan, write, discuss, tear up, revise, and write again and again. As teachers, we hope that in addition to learning skills, students somehow will learn to love a writing process that allows them to discover something of themselves and the world around them as they think through problems and learn to communicate their ideas in effective prose.

Our colleges and universities must decide what they value and what skills their students must have before they develop testing rituals. Each institution must weigh the benefits and disadvantages of different models of testing. Short-answer tests may have economic and temporal advantages, but they have gross disadvantages: they cannot assess the important rhetorical skills that students must learn, and they cannot elicit the kind of writing that our literate community professes to value.

Whatever our reasons for testing writing, the instruments that we develop will be, of necessity, imperfect. Whether we test for competence or excellence, to sort or to rank, we borrow, knowingly or unknowingly, methods used 1300 years ago to evaluate writing and thinking. And we suffer from the limitations of whatever assessment instruments we choose—as did the Chinese in centuries past. We agonize about the possibility that our tests will discriminate against students who have not had adequate preparation prior to the time we test; we worry about reader bias in essay testing; and we argue about the long-term effects of our tests on our students' writing. Is form dominating content and stifling creativity, as the Chinese feared in their "eight-legged essay"?

Ritual and testing are interrelated, as we can see in the initiation rites of early societies. The values of a group are symbolized in the tests one must pass in order to become a member of that group. We are being forced to test outside of college courses today because as educators we have refused to agree on and articulate our values within our courses. That there is a general distrust of college faculty is exemplified in the statewide and citywide involvement in testing in colleges and universities. Early societies developed *rites de passage* that reflected their values and their needs, depending on the way in which they lived, worked, and believed. Within the group, admission into adulthood depended on the ability to demonstrate mastery of specific tasks. So we in colleges and universities today must decide on the values and needs of membership in the group to which our students aspire. If they need skills in thinking and in making connections between disparate ideas, if drawing material together into a coherent written whole is vital to membership in a group of educated adults, essay tests will be part of our essential rituals.

As faculty and writing program administrators, we must assume leadership in assessment. We must clarify and profess our values. What do we want our students to know? What kind of thinkers should they be? What will they need to move into the complexities of the next century? Our tests should be *rites de passage* to help our students live well in that world.

REFERENCES

China Handbook Editorial Committee. *Education and Science.* Trans. Zhou Yicheng, Cai Guanping, and Liu Huzhang. Beijing: Foreign Languages Press, 1983.

Cramer, Richard. "Testing Multiple Study Choices." *Psychology Today* May 1984: 17.

DuBois, Philip H. *A History of Psychological Testing.* Boston: Allyn & Bacon, 1970.

Ebel, Robert L. *Essentials of Educational Measurement.* Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Fairbank, John K. *Chinabound: A Fifty Year Memoir.* New York: Harper & Row, 1982.

Fairbank, John K., Edwin O. Reischauer, and Albert M. Craig. *East Asia: Tradition and Transformation.* Boston: Houghton Mifflin, 1973.

Goodlad, John I. *A Place Called School: Prospects for the Future.* New York: McGraw-Hill, 1983.

Heren, Louis. *China's Three Thousand Years: The Story of a Great Civilization.* New York: Macmillan, 1974.

Miyazaki, Ichisada. *China's Examination Hell: The Civil Service Examinations of Imperial China.* New Haven: Yale U, 1981.

Rank, Otto. *The Myth of the Birth of the Hero and Other Writings.* New York: Random House (Vintage Books), 1964.

Scharfstein, Ben-Ami. *The Mind of China.* New York: Basic Books, 1974.

I am indebted to Mr. Kuang-fu Chu, Chinese specialist of the Oriental Division, New York Public Library, for generously agreeing to review this manuscript. His experience and enthusiasm were of enormous help.