Language Arts Journal of Michigan

Volume 10
Issue 1 *Focus: Tough Questions. Hard Choices.*

Article 5

1994

Reflections on Implementing Writing Assessments

Ronald A. Sudol

Reflections on Implementing Writing Assessments

Ronald A. Sudol

Assessment is so intimately bound to our conceptions of teaching and learning that it's impossible to imagine a system of public education without it. The onset of mandated large-scale assessment may make us wish we *could* get by without it, but the tough questions and hard choices we face necessarily have more to do with accommodation than avoidance. The *Michigan Proficiency Examination Framework for Writing*, published in March, 1993, rests on solid theoretical ground, but its effective implementation depends on continuing informed discussion about the aims and methods of writing assessment. In her review of the development of the *Framework*, Ellen H. Brinkley credits the influence of the many teachers who participated in the advisory councils and site meetings conducted by the framework management team: "Too often . . . English language arts teachers are inclined to assume that others are the experts. . . . We realized as we worked through our long sessions that the perspective and effort of every one of us involved was needed if we were to make a positive difference" (34). Such empowerment is welcome, to be sure, but it also demands that we negotiate competing perspectives on assessment with clear-headed diligence.

The problems of implementation fall into four categories. First, there's the problem of inconsistent and sometimes conflicting definitions of assessment among teachers, administrators, psychometricians, and lawyers; second, there's the problem of defining who is really being assessed and why; third, there's the problem of the tricky relationship between assessment and curriculum; and fourth, there's the curious but generally understandable resistance against state-wide assessment by teachers.

"The curriculum frames the assessment, but the assessment, in turn, drives the curriculum. This reciprocity is so fundamentally correct and deeply felt that it can sometimes obscure another fundamental notion—that assessment and curriculum are and must be different things."

The first problem is definitional. The Latin root of "assess" is "to sit by," suggesting to the sensibilities of educators a kind of tutorial intimacy where testing and teaching enjoy an easy and recursive relationship. But we should not miss the irony in the fact that both ancient and contemporary usage connect assessment to taxation, and it is the public tax burden that has fired the engines of mandated assessment in schools. Indeed, the word resonates with images of the kind of precise measurement, detailed record-

making, and due process we associate with the proper functions of government.

Psychometricians enjoy high status in bureaucracies and courts because their peculiar expertise comports well with these public notions of assessment. The gulf separating the teacher's view of assessment as tutorial and developmental and the psychometrician's as measured standardization is not a simple difference of opinion. It is a difference between separate intellectual domains defined by different world views, assumptions, and systems of logic. What the teacher may see as a perverse numbers game, the psychometrician sees as fairness and truth, pure and simple. We should understand the scale of these differences when we argue, compromise, or dig in our heels on any assessment issue.

For example, an earlier version of the extended standard task in the Michigan *Framework* provided for peer conferencing and editing, on the assumption that the assessment should mirror exemplary classroom practices. The reality of assessment as measurement, however, disciplined us into either abandoning or modifying these features because they introduce inconsistency and uncertainty into the assessment. The variations in student performance that make collaborative learning work for teachers in classrooms are, in the psychometric domain, examples of unacceptable variables likely to contaminate the test results. Such variables become, in this domain, a lapse of fairness and equal opportunity and might very well be actionable. It's not hard to imagine what a lawyer could do for a client who failed a writing test because he or she did not have access to the same quality of peer conferencing that other students enjoyed.

On the other hand, we did not compromise on the matter of so-called objective testing. It is easy to argue that the only valid test of writing is *writing* and not answers on a multiple choice test. Nevertheless, many writing assessments combine the scores of multiple choice tests with the scores for writing samples in an effort to balance the high validity of writing samples with the high reliability of mechanized scoring. We felt, however, that the practice of scoring writing samples holistically has been developed and refined to the point where trained human judgment can be reliable and consistent.

Still, psychometricians press one additional point: that the consistency of the writing assessment from year to year needs to be verified by administering a parallel standardized test. However, to use an *indirect* measure (that is, a standardized multiple-choice test) in order to validate a *direct* measure (that is, an essay written over an extended period of time) is illogical, and the flaw is demonstrable entirely within the domain of measurements professionals. Exposing the flaw, however, does not solve the problem, one of several still pending. Holistic scoring works best in the closed environment of a single batch of samples. Comparing the results of one year's writing with another year's is thus problematic. How can we assure that the assessment be equally difficult and the scoring equally rigorous at each session? We can do it by creating a cadre of trained holistic readers among Michigan teachers and by designing essay topics and scoring rubrics accordingly.

"A good writing curriculum, under current orthodoxy, will be process-oriented, but a good assessment can be nothing but product-oriented."

So, in balancing these sometimes conflicting views of assessment, the Michigan framework is exemplary. It does not include any machine-scoreable items whatsoever; the main writing task extends over two days; and it provides for crediting of student-selected portfolio pieces of writing. Tests and measurements professionals have shown a respectable amount of courage and imagination in going along with these departures from what they would normally consider to be standard practice.

The second problem has to do with who is being assessed and why. The problem is more political than pedagogical, and there is hardly space here to deconstruct the agendas and ideologies underlying statewide assessment. In the absence of directives to the contrary, the manage-

ment team proceeded with pure hearts and tied the assessment to a curricular framework. As we get closer to the first administration of the test, however, the other agendas will start asserting themselves.

One of these has to do with how the reporting of scores affects public perceptions. The reporting of scores for the existing Michigan Educational Assessment Program, in which various subjects are tested at various grade levels, has caused quite a bit of consternation. Scores are presented compositely by school district and reported in the press as measures of the quality of teaching in one district as compared with others. Articles and graphs clipped from newspapers are sometimes posted in real estate offices to help home buyers decide where to live. Inferring school quality from these composite scores undermines the entire enterprise by recklessly entangling assessment scores with demographic information.

Even though the reporting of scores and their interpretation by the public are largely beyond the control of public education professionals, these external factors obviously impinge on the quality and effectiveness of the assessment itself. Given this context, our deliberations about how many samples of writing we need to measure; whether or not this should be a pass/fail test of minimum competency; or whether we should provide for a full range of scores, including those representing distinguished performance, are likely to generate some heat. Another example of a political entanglement is the relationship between assessment results and teacher and school accountability. Clearly, settling these non-curricula issues ahead of time makes implementing the assessment easier, and it should make the results more useful.

The third problem has to do with the tricky relationship between assessment and curriculum. It is a truism, of course, that assessment and curriculum must have a reciprocal relationship. The curriculum frames the assessment, but the assessment, in turn, drives the curriculum. This reciprocity is so fundamentally correct and deeply felt that it can sometimes obscure another fundamental notion—that assessment and curriculum are and must be different things.

A good writing curriculum, under current orthodoxy, will be process-oriented, but a good assessment can be nothing but product-oriented. We simply cannot do more than assess a particular exhibit—the end product of an individual performance that may reflect any of thousands of different processes. Our enthusiasm to keep assessment from lapsing into nothing but numbers (and the woeful things that can be done with those numbers) can seduce us into trying to assess processes instead of products, to blur the distinction between assessment and curriculum. It's bad enough when the writing process gets broken down into discrete stages, worse if all students get shunted like a herd through these stages, and worst of all if we try to assess a student's mastery over any element of the process in any kind of criteria-driven way.

"We like to think that brainstorming, multiple drafting, conferencing, and revising help students write better. Probably they do. But in the end, only the writing matters."

We like to think that brainstorming, multiple drafting, conferencing, and revising help students write better. Probably they do. But in the end, only the writing matters. In any case, the familiar list of items that constitute the writing process are only the most visible elements of an activity that is largely invisible, mysterious, various, and unstable. The management team abandoned all of its earlier attempts to specifically incorporate "process" in the assessment. "Process" exists in the curriculum and theoretical background of the framework—in recognizing writing-across-the-curriculum, in encouraging metacognitive reflection, and in allowing ample time for incubation and editing. The challenge for teachers is to find ways to help students internalize their individual writing processes. The assessment instruments themselves cannot be expected to assure students will perform the preliminary and revisionary activities that might improve their performance.

The most important link between the curriculum and the assessment is having local teachers evaluate writing samples using the procedures of holistic scoring. The challenge here is to overcome the assumption that such a process is subjective and inconsistent. Anyone who conducts training sessions in holistic scoring knows that any group of English teachers, no matter how competent, will score the same set of essays with wide variations until they have gone through a well designed program of consensus-building. This is not to impugn their competence but simply to recognize that grading essays in the classroom and scoring writing samples on a high-stakes assessment are very different activities, requiring different approaches, different criteria, and different forms of accountability.

Typically, the holistic scoring of writing samples is validated by a parallel multiple choice exam. One expects to see a correlation between the human and computer generated scores. The logic of this procedure is just as faulty as the logic used to justify machine-graded tests to validate the assessment from year to year. Since the *Framework* makes no provision for machine-graded testing, the holistic scoring must come as close to perfection as possible—and this will certainly be a challenge. Those who read and score writing samples will need to set aside their individual criteria, work toward building consensus about what to value in writing, and fairly and consistently apply that consensus to all 110,000 essays year after year. This is a tough job, and doing it well would reflect a high level of professionalism.

The challenge of scoring essays fairly and consistently raises a question about the term "assessment" itself. This word is now so widely accepted that there is probably no replacing it. But if we could, I wonder if "appraisal" might not be better name for this activity. "Appraisal" seems less quantitative. The appraisal of real estate, for example, with its use of human judgment, lists of criteria, models, and multiple measures shares important features with holistic scoring. In both cases one tries to put a value on what is there. Appraisers assume responsibility for their judgments, and their work is monitored by equally qualified colleagues. Moreover, the word sounds better. "Assessment" has become a hissing expletive.

The fourth problem is resistance to assessment itself. At the site meetings throughout the state teachers would often preface their otherwise positive remarks by going on record as being opposed to statewide assessment. They say they are willing to go along only because doing so seems the shrewder course of action. But, they say, they have detected the hidden agendas; they have been battered by all kinds of public slanders and outrages; and they see more and more precious classroom time devoted to state-mandated activities.

"The most important link between the curriculum and the assessment is having local teachers evaluate writing samples using the procedures of holistic scoring. The challenge here is to overcome the assumption that such a process is subjective and inconsistent."

True enough. But teachers are already deeply involved in the assessment business. Being able to assess student performance against recognized standards is a necessary prerequisite to effective teaching. Necessary but not easy. Switching between the roles of teacher and judge is one of the most stressful things we are called upon to do. Peter Elbow has written persuasively that separating these roles can relieve this stress and unleash creative teaching. A well-designed statewide assessment can serve that function. And let's not forget that writing, even under duress, is epistemic, an activity from which we should always learn. No other school subject can claim its assessment has as much potential to be a learning experience.

When the time comes to prepare to score the assessment samples, statewide training in holistic scoring can become a potent form of staff development. In addition, sharing scoring rubrics with students and having them score sample essays is a highly effective collaborative class-

room activity. Training teachers of writing does not begin with theory. It begins with assessment: What do people actually write? Why do they write that way? What makes one piece of writing "better" than another? Answering such assessment-type questions leads to theory, then to practice. So one way to deal with the frustrations of mandated writing assessment is to reaffirm its position as a central element in professional growth.

"Being able to assess student performance against recognized standards is a necessary prerequisite to effective teaching. Necessary but not easy. Switching between the roles of teacher and judge is one of the most stressful things we are called upon to do."

Finally, language arts professionals should recognize that the assessment plan itself is developmental and that they have an important role to play in its evolution. Their roles in holistic scoring, setting standards, and reforming curriculum have already been mentioned. In addition, however, the framework document really is a *framework* in the sense that it establishes categories of assessment that can be accomplished by different kinds of tasks. The two samples of outside writing called for in Strand I set the stage for continued development of writing-across-the-curriculum programs in the short term, but in the long term, this strand opens the door for portfolio assessment whenever we find a reliable way to score portfolios. The writing in the other two strands can be done with computers whenever there are enough computers to enable all students equal access. The generous time allowance in Strand III permits innovative approaches toward designing writing prompts, and the elimination of a severe time constraint should allow progressively more stringent expectations of writing and editing performance. Thus, the *Framework* enables the assessment and the curriculum to grow together.

Works Cited

Brinkley, Ellen H. "Responding to Michigan's Legislative Mandates." *Language Arts Journal of Michigan* 9.2 (1993): 31-34.

Elbow, Peter. *Embracing Contraries: Explorations in Learning and Teaching*. New York: Oxford UP, 1986.

Michigan Proficiency Examination Framework for Writing. Michigan Council of Teachers of English. 31, March 1993.