Rexford Brown

WHAT WE KNOW NOW AND HOW WE COULD KNOW MORE ABOUT WRITING ABILITY IN AMERICA

In the area of measurement of growth and proficiency in writing, one of the major difficulties has always been that word "measurement." Few of us in the English teaching profession feel comfortable with the associations of precision and icy objectivity that accompany the word "measurement," and most of us were brought up thinking you can't "measure" writing. All of us have been "grading" essays for years—by which I mean doing a range of things from simply saying "uh huh" to students as we hand back their virginal papers, to actually granting two or three letter grades and obliterating their text with such strange glyphs as "awk", "punc", "frag", "dang", and "rewrite by Friday."

A major advantage to the word "grading" seems to be that it supports the widespread feeling among too many of us that standards for evaluation of writing are somewhat personal. We are all very careful to respect each other's right to a private grading system, even if it is arbitrary, wrong-headed, nasty, or capricious. Criticizing a colleague's values in this area is academically equivalent to crossing a picket line. In such an atomistic climate there has been little room for the idea of measurement because we have assured ourselves that there are no shared units of quality, there is no bureau of standards. Proficiency in writing in this climate is expressed as a letter grade, and growth can only be expressed as an improvement in grade. The fact that a writer can improve his writing between his freshman and sophomore years, but receive a lower grade because his second teacher holds different views from the first, bothers no one but the poor student. Nor do we seem particularly concerned about the fact, easily borne out in a number of studies, that our own grades are subject to many kinds of bias, and fluctuate randomly in ways that few of us can control.

Rexford Brown is Director of Publications with the National Assessment of Educational Progress.

DOI: 10.37514/JBW-J.1978.1.4.01

It is interesting to note that the first major advance toward large-scale measurement of writing samples was successful largely because it did not seriously threaten the picket-line principle or challenge in any way the professional conspiracy of silence about quality. Educational Testing Service has for many years called together radically different people, trained them to recognize certain papers as 3s or 6es (don't ask why, just internalize the training papers), lavishly paid them to read hundreds of papers and respond knee-jerkily to each with appropriate scale numbers—and never ask any of them to lose face by revealing that they might have harbored perverse or insane notions about what constitutes quality writing. Holistic scorers need never explain what they are doing; and thus did holistic scoring achieve a certain amount of respect in our profession. Measurement got a foot in the door by pretending it was not measurement.

We've learned that large numbers of essays can be reliably scored with the holistic method and that these scores are accurate predictors of college success. And we've learned that teachers can be trained to agree on something. But what do holistic scores mean? All anyone knows after a holistic scoring is that paper A is higher on the scale than paper B. But, since no one discussed quality criteria, no one knows why. Furthermore, it is possible that all of the papers at the top of the score are horribly written. They may be better than the rest, but still may be unacceptable to most teachers of composition.

Not only is this traditional holistic scoring incapable of establishing proficiency in any concrete sense, it is a very unsatisfactory system for the evaluation of growth. If a student's first paper is rated 5 at a September scoring session and her 20th paper is rated 6 in a May session, we know nothing, because experience has shown that holistic scorings cannot be replicated reliably. We know more about growth if both papers are included in the same scoring session and the second paper comes out higher on the scale; but we still don't know why it is better or how good it is in an absolute sense.

No matter how reliable holistic scoring is as a way of rank-ordering papers it is inadequate as a measuring tool in itself because it is entirely relativist and value-free. It is not tied to any absolute definition of quality. The most promising *modified* holistic scoring approach I know of is the National Assessment of Educational Progress (NAEP) "Primary Traits" system. Developed to counter a glaring fault in traditional holistic scoring—that you cannot report results in useful or even meaningful ways—the system rests upon elementary rhetorical

theory. It assumes that a carefully defined writing task is a statement of certain rhetorical imperatives; that successful completion of the task entails understanding of and responsiveness to those imperatives; and that degrees of success are definable in concrete terms. We have found the tasks hard, but not impossible to define, the scoring guides complicated but teachable and the actual scoring reliable. Most importantly, we have found the results reportable in terms that have curricular implications.

For many teachers, holistic scoring has been a luxury only the rich could afford anyway. Still reluctant to define quality, but nevertheless in need of evaluation systems, they have used objective, multiple-choice tests of writing ability. Such tests are cheaper and easier to score; best of all, they enable any user to say "Well, *I* sure don't define writing the way those test developers do, but I'll accept their claim that the results correlate with writing ability; and after all, these are the only tools available."

But machine-scorable tests *also* suffer from some glaring weaknesses. Their primary function is, again, to rank order people on a scale. This leaves us again with no absolute knowledge about writing ability and a slight sense of embarrassment when we tell people we'll test their writing ability by not requiring them to write a single word. *Of course* these tests correlate with writing ability and predict academic success; but the number of cars or television sets or bathrooms in one's family also correlate with his writing ability, and parental education is one of the best predictors there is. All existing objective tests of "writing" are very similar to I.Q. tests; even the very best of them test only reading, proofreading, editing, logic and guessing skills. They cannot distinguish between proofreading errors and process errors, reading problems and scribal stutter, failure to consider audience or lack of interest in materials manufactured by someone else. Like holistic essay scoring, multiple-choice testing of writing is seldom diagnostic in any useful way. And since capacity to recognize problems in other people's writing does not insure capacity to avoid them in one's own writing—especially first draft writing—we can never be sure what the final scores on such tests mean, let alone the subscores.

There are even more insidious aspects to multiple-choice writing tests. They require a passive, reactive mental state when actual writing requires and fosters a sense of human agency, an active state. And they are necessarily incomplete, leading the student and perhaps even the teacher to believe that those aspects of writing most easily tested—sentence

structure, word meaning, spelling, punctuation and outlining—are the most important to teach and learn. Finally, since the approach of many such tests is to emphasize differences between standard and nonstandard usages, writing courses all too often become, unintentionally, cultural programming laboratories.

No, an objective test all by itself is not a very good measuring device either; it tells us something, but not enough that is concrete. But the proliferation of such tests over the years has softened the profession up just a bit more toward the idea of measurement and the possibility that there are some shared units of quality upon which to build more accurate and useful systems of evaluation.

We're ready now to work toward the creation of many such systems. The pressure is on from the public, the deans, and the students themselves to improve writing. In order to do it, we're going to have to know more about the process of composition than we do now, and we're going to have to know more about what is wrong—in concrete, absolute terms—with student writing. Even our agelong system of medieval fiefdoms—separating the Miltonians from the linguists from the English educators from the modernists from the rhetoricians from the Marxists from the graduate-student assistants who teach freshmen composition—even that is crumbling under the economic and social pressures so familiar to us all; and this crumbling makes possible a movement toward professional discussion of quality in writing. The picket-line principle is doomed.

We have learned a great deal in the last fifteen years about the strength and limitations of the various holistic scoring systems developed at ETS, National Assessment and elsewhere; we know what is useful and valid in such good objective tests as the Houghton Mifflin College English Placement Test and the ETS STEP test; our knowledge of syntactic maturity levels has been advanced by the work of people like Walter Loban, Kellogg Hunt, Lou LaBrant, Roy O'Donnell and others; the contributions of John Mellon and Frank O'Hare to our knowledge of the relationship between sentence combining activities and syntactic maturity levels have opened new and exciting evaluation opportunities; the rebirth of rhetoric, and the particular contributions of Francis Christensen, Ross Winterowd and Edward Corbett have given us new frames of reference for definitions of quality that facilitate concrete evaluation.

We can create from this fund of knowledge and this special climate a number of evaluation systems that define proficiency in concrete terms, are sensitive to degrees of growth toward that proficiency, require people

both to write essays and test their editing skills, are valid and reliable, are cheap and—most importantly—*are coordinated with the long-range research effort we need to more fully understand and develop strategies for improving the process of composition.*

Here are some suggestions about how to develop an ideal instrument:

1. Make students write—but there's no need for more than 400 words on test essays.
2. Base essay evaluations on papers reflecting several models of discourse, because quality differs for each one and people are not equally proficient in all of them.
3. Teach testers how to write directions for essay examinations. If you want to evaluate an essay for certain characteristics, then you must be sure that you have requested them in the assignment. This is not a trivial matter: it is extremely difficult to write assignments that define precisely the rhetorical imperatives that will either be met or missed by the students. If you want to know whether they can elaborate upon a role expressively while maintaining control of point of view and tense then you have to set the task up in such a way that they must do so, and define acceptable levels of achievement that are concrete and realistic.
4. Use computers. Have people mark off T-units in the essays so you can gather information about number of words per T-unit, number of clauses per T-unit, number of words per clause, number of adjective clauses, number of noun clauses, and so on— information about embedding, in short, which ties you directly to indices of syntactic maturity.
5. When you have these counts, tie them to holistic scores. If the scorers cannot or will not tell you why some papers are better than others, the computer will at least give you an idea of what was influencing them.
6. Tie the counts to various criterion-scoring systems. The six factors that seem to affect judgment most are ideas, mechanics, organization, vocabulary, what Paul Diederich calls flavor, and handwriting. Each can be evaluated independently.
7. Define coherence in specific syntactical or transformational terms, have graduate students code papers accordingly and establish a concrete coherence scale.

8. Include in any instrument questions about writing attitudes, prewriting activities and rewriting activites and then look at results in the light of that information.
9. Require basic sentence combining exercises and tie results on such exercises to actual writing performance.
10. Include a battery of objective items that will at least remind students that they should edit.
11. Use the resources we already have. National Assessment's huge corpus of essays remains largely untouched by researchers. Ross Winterowd at the University of Southern California has received seed money from the NCTE and Carnegie Foundations to keypunch representative samples of NAEP essays for research into the syntactic features of coherence and other vital matters. But the research undertaking itself has not yet been funded. Various graduate students here and there have used bits and pieces of the corpus for various projects, but they have only scratched the surface. Much that could enrich our understanding of the composing process and those aspects of it that cause most confusion for students of writing remains undone. The situation will probably continue until more is known about the availability of national and state assessment score materials.

The next national assessment, currently under development, will include most of these features. In addition, it will include materials from 1969, enabling us to examine trends spanning a decade. But however good it is, it will not be sufficient to gather all the information we need at the necessary level of detail. For that we need a coordinated effort involving writers, teachers, linguists, anthropologists, rhetoricians, philosophers, data gatherers, and educational psychologists. Professional conferences, which bring together such people, must serve as the model for the inter-disciplinary approach which alone can promise sufficiently sophisticated understanding of our situation. Perhaps, after more such meetings I will be able to provide more concrete information about achievement in writing and more exciting and practical specifications for its assessment.