HOLISTIC ASSESSMENT AND PROFICIENCY TESTING

Proficiency testing is increasing, especially in California where it is now required by law. In most school districts the tests are designed by local teachers and administrators. Thanks to a growing concern about the effects of testing on what is taught, more and more schools are including writing samples in their overall assessments, requiring some training for teachers in the holistic scoring of essays. And holistic assessment is the one part of proficiency testing which contributes directly to staff and curriculum development: teachers gain an important perspective on what they are accomplishing as they read papers representing a wide range of student ability.

But it is one thing to collect writing samples when we are interested in how a group performs, and quite another to collect samples which will determine whether an individual should be granted a diploma or admitted to the college of his choice. The responsibilities of the tester increase proportionally to the importance of the outcome for individual students.

Whether they are evaluating groups or individuals, test designers must consider several critical factors. The importance of a good topic cannot be overstated (see article page six), nor the importance of collecting more than one sample of writing from each student, nor the importance of allowing adequate time for composing. But there are further considerations which are necessary to ensure that holistic assessment in proficiency testing will yield fair and accurate outcomes. The suggestions made here are based on the most conservative current practices, with the understanding that new research may eventually establish other procedures which are equally valid or even superior.

MAKE PASS/FAIL DECISIONS AFTER PAPERS ARE SCORED

The best time to set the “passing” score is after the reading process. Teachers and administrators who are charged with establishing the proficiency cut-off point may arrive at a far more realistic and defensible point if they have read papers from a wide range of student performances while hearing the responses of other readers. These decision makers will then be in a better position to explain their decisions to students who do not pass and to defend their decisions to parents.

Of course, the main components of “proficiency”, as it is defined in a given district, should be made clear before the test—to students as well as teachers—to insure fairness and maximum benefit from preparing for the test.

The criteria for determining proficiency should also play a part in developing the scoring guide and training readers. The scoring guide should not, however, be transformed into an analytic count of certain qualities. A primary advantage of holistic scoring is that it allows us to register the total effect of a piece even when that effect cannot be described by naming particular qualities. Furthermore, not all of the scoring will be based on the concept of proficiency or minimum competency. The range of scores will allow distinctions among different kinds of proficient papers—good and excellent, for example. Lower scores may discriminate between totally non-literate and barely literate writers.

The criteria for defining proficiency, together with the characteristics of each level of quality which are set forth in the scoring guide, are important both in determining cut-off points and in interpreting scores to the public. In one district, for instance, teachers may find that scores of 3 have been assigned to papers that are, by their standards, clearly non-proficient, with scores of 4 being assigned to papers generally proficient. Other districts may discover that only papers receiving scores of 5 or 6 can be called proficient by their standards. Such differences can be expected, not only where standards may differ, but where student populations vary dramatically in level of achievement. Establishing consistent definitions of proficiency and comparable scoring guides from one assessment to another remains a problem to be faced in the near future by political as well as educational communities.

USE SEVERAL DIFFERENT KINDS OF MEASURES.

Objective measures of composing skill cannot adequately substitute for samples of student writing, but such measures may, if carefully designed, yield information which is less economically obtained from writing samples—reading comprehension, vocabulary recognition, editorial skills, for example. Objective tests of general language competency help increase the reliability of a student’s overall score because these tests are scored exactly the same way each time. Also, many teachers are more comfortable reading holistically if they know that someone, somewhere, is “counting errors.”
Under the system advocated here, it should not be possible for teachers to boast, as one did recently, "I didn't give a single 6." This teacher wanted no one to doubt his high standards for excellence. The assigning of pass/fail significance to scores after the reading can help wean teachers from defensive or punitive uses of scores because the emphasis is not on grading papers, but merely on rank-ordering them. Finally, this procedure ensures that even the best and most reliable readers will not be affected in unpredictable ways (as they sometimes are) by knowing that a particular score means pass or fail.

ESTABLISH A BORDERLINE RANGE OF SCORES

After the holistic scoring the teachers can sort students into three groups: the clearly proficient, the clearly incompetent, and those who need further diagnosis. Keep in mind that both writers and readers are to some degree unreliable, even though a well-designed assessment will minimize this unreliability. The combined ratings of two readers are, of course, more reliable than one rating, but the student himself may not have given a representative performance. The margin of error in holistic assessment, which usually runs 10% or more even under the best conditions, suggests a second look at students who score close to the pass/fail line.*

Again on a six-point scale, if a "competent" paper is defined as four or better, the uncertain range should include:

<table>
<thead>
<tr>
<th>1st and 2nd readers: 3 + 3</th>
<th>3 + 4</th>
<th>4 + 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total score:</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

It is just as likely that students on the lower edge of passing will fail as it is likely that students on the upper edge of failing will pass a second test. Passing seniors, may perhaps be given the benefit of the doubt, but lower classmen who barely pass may need re-testing to establish their competency unambiguously.

FALL-BACK MEASURES

The "second look" at students in the borderline range might consist simply of a re-reading of their essays. If the additional reading puts a student into the clearly proficient category, the problem is solved. But any unfavorable re-reading should invite another measure of the student's writing performance. A second testing, or a portfolio of classwork, for example, may reveal the proficiency necessary for functioning in the real world.

Opportunity for reassessment should be made available to any student, on appeal by the student and at least one teacher who believes the test score does not reflect the student's actual ability.

CONSIDER WHAT PASSING MEANS

When passing the writing test exempts a ninth or tenth grader from future English courses, the tendency is to set the pass mark quite high to prevent too early withdrawal from the English program. Students need the practice and development that comes from further work in language if they are to maintain proficiency or perform well at increasingly sophisticated language tasks.

But a high passing level designed to keep students in the program is unfair to those who have completed the program with little more than minimum competency after years of work and concentrated help from teachers. Passing levels should be based on a careful definition of minimum competency. At the same time, a measurement of several levels of proficiency above minimum competency provides a useful guide for determining the needs of students in the English program.

MAKE PROVISIONS FOR DIAGNOSIS AND REMEDIATION

A good way to use assessment results is to allow failing students to discuss their essays with a teacher who can help them diagnose their strengths and weaknesses as writers. Groups of students might read and re-score sample essays from the assessment to learn the expectations of the audience which rates them, while observing other students' strategies for responding to the test.

Diagnosing students who need special help is a major justification for proficiency testing. But in some schools the problem does not seem limited to a few students. When a large percentage of a school population "fails" the writing test, an overhaul of the entire program may be indicated, and the assessment results are a healthy shock.

But in some cases, the assessment itself may be called into question. Were pass/fail scores arbitrarily set before the sample was collected? Were definitions of competency set arbitrarily high to keep students from thinking their educations were complete? Were they based on a realistic understanding of functional literacy? Were teachers defending themselves from charges of "lowering standards" by setting standards impossibly high?

CONCLUSION

With some attention to proper assessment methods and appropriate use of scores, we can continue to demand funding for this valuable examination of our students, our program, and
Holistic Assessment and Proficiency Testing

ourselves. A great benefit of assessment is that it helps to create cohesion, coherence, and continuity within the community of teachers who participate in it. It helps to refocus the blurred vision and broaden the narrow perspectives we so often develop when working in isolation. But we cannot defend this benefit to ourselves, however much our teaching is improved, if at the same time the rights of individual students are not rigorously protected.

With the widespread use of holistic assessment by local schools, good and bad adaptations of the model will inevitably emerge, some in response to imaginative insights into the writing process, some in response to economic belt tightening, some in response to defensiveness about "our standards" or "our program." In coming years there will be increasing need for Writing Project leadership in refining and stabilizing the methods of holistic assessment at the local level if we are to maintain the position that this measure can be reliable and valid, fair and meaningful to students, teachers and parents alike.

*Footnote. There is as yet no generally agreed on method of calculating the standard error of estimate for holistic assessments. This estimate is based on a number of reports of reader reliability and reports of research on writer dependability. For calculating actual error of holistic scores, the following are useful:

Diederich, Paul, Measuring Growth in English, NCTE, 1974.


Catharine Keech is a Research Assistant with the Evaluation Unit of the Bay Area Writing Project.