of the placement program. At Pittsburgh, Smith tried three different assignments: (1) Students were given an open topic; they were asked to write about a time when they were creative. (2) Students read a passage by D. H. Lawrence that described a time when he was creative. They were then asked to write about how Lawrence was creative and asked to respond in terms of their own creativity. (3) Students were given three passages by separate authors (all about being creative) and asked to read and respond to them. On the first assignment, the high-scoring students were rated far superior to the middle and low writers. On the second assignment, the high, middle, and low-scoring students could still be distinguished, but their scores were much closer. The third assignment, which required the reading of three passages, easily confused the best and the worst writers, while both the average and superior writers scored higher and closer to each other. Such results prompted Smith to point out that a school needs to try out various topics and discover which ones suit the students best. She concluded that the department actually wants to consider. Smith suggested that teachers adapt topics that have proved not successful in the classroom. Such topics should be fairly representative of what a student will actually encounter during the course of the semester in order to better predict classroom performance.

In determining how many topics to use, one must consider how much time and money are available for placement programs. If one uses three twenty-minute topics, the raters will need more time, but they can get a fairly good idea of what a writer's average work is like. Once again, one must decide whether it is an average writing sample that the program elicits or whether it is the students' best writing or worst writing that is desired. When evaluating the essays, raters must decide what it is they are to use as criteria for their evaluation—production of words, gut feelings, coherency, or something else. In any case, their criteria must be stated so that someone else can use them to help the students.

The actual conditions under which students take a placement test, of course, make a difference in their performance. After extensive evaluation, it appears that students do best on a summer placement test if it seems as much like a classroom test as possible. When given two hours to complete their writing, 74 percent of the students in Smith's study took between sixty and ninety minutes. Only 6 percent took more than ninety minutes. However, those who took less than sixty minutes actually wrote more than the other students. Other findings were that smaller testing groups produced better writing, as did the presence of a teacher and the use of a warm-up activity. In determining the reliability of the raters, Smith and his colleagues followed up on their students. They found that students who were simultaneously rated very high and very low by different raters tended to get better grades as they progressed in their writing courses than did those who were consistently rated as "middling" by various raters. As a result, this particular phenomenon of wide-ranging scores is also taken into consideration when placing students.

Smith concluded by reminding the audience not to buy anyone else's system, for if it doesn't fit, a school may be buying someone else's mistakes. He encouraged each institution to consider its particular needs, time, money, and faculty commitment. If an assessment program is theoretically sound and its directors can sleep at night, it will probably be quite satisfactory, given enough time, patience, and imaginative hard work.

---

**MATCHING ASSESSMENT PURPOSES AND INSTRUMENTS**

Speakers: Gretchen Glick, McGraw-Hill, Inc.
Barbara Cole, McGraw-Hill, Inc.
Introductor/Recorder: Ellen Kilgo Summy, University of Akron, Ohio

Gretchen Glick discussed four interrelated options that testers and educators must consider in order to match a writing test to the testers' purposes: formats, prompts, scoring, and reporting. When choosing a test, testers must first consider the information they want to obtain, the purpose of the testing, and their resources—both financial and temporal. Indirect assessment of writing, done through an objective instrument, may be adequate and desirable in some circumstances, but most programs use direct measures (a writing sample) for purposes of placement and admissions, research, and proficiency. The ultimate purpose of direct writing assessment is, of course, to improve instruction and to benefit students.

Once the testing purpose is established, testers must develop an assignment or prompt designed to produce the kind of information that they want. Mechanical skills, such as punctuation, spelling, and vocabulary, can be evaluated from a diversity of prompts. However, if the purpose is to evaluate content or organization, all students must be given one prompt, and that prompt must have directions about the topic and the type of writing, whether descriptive, narrative, expository, or persuasive. Then, all papers written in response to a prompt can be compared to a standard—either a model or a set of guidelines. When all students write to the same prompt, their papers can be compared to each other and ranked. In choosing the prompt to be used, testers must consider their program, its texts, and its writing instruction. They should also consider what type of writing is appropriate for the students' ages and abilities. Finally, the topic used should interest students, and the prompt should be general enough to give them something to think about, but specific enough to focus the content of their writing. Sometimes, more than one prompt will be used to provide more information to evaluators; however, the testers' rule is to do the minimum amount of testing that will provide the information needed.

Next, testers must consider scoring options, including the method, the scale, and the location. Although there is no nationwide consensus on the best method for scoring writing samples, the three used most are holistic, analytic, and primary-trait (but even these labels have different meanings to different groups of educators). In choosing a method of scoring, testers must focus on two issues: the testing purpose and the kind of information desired. For instance, holistic scores can be used to evaluate programs and to rank students; analytic scoring can give teachers diagnostic information on individuals' weaknesses so that they can plan classroom instruction. The point scale for rating is also based on the testing purpose. To separate writers into ability levels, a four-point scale might be used: good, acceptable, below average, unacceptable. But to give a more detailed evaluation, an eight- or ten-point scale might be used. The choice of a location for scoring involves the choice of evaluators. With purpose as well as resources in mind, testers could choose professional evaluation with experienced evaluators and computer-monitored, bias-free scoring.
Or they could choose local scoring that brings teachers together to discuss their writing programs and their expectations; these local evaluators would, then, have to be trained as readers.

In reporting testing results, testers must tailor the form of the report to its audience. For example, a summary report will go to administrators at different levels who want to know how groups of students, or schools, or districts compared to each other. Classroom teachers need information on individuals; perhaps holistic data for grouping by general writing ability and analytic data for determining specific instruction. And parents will want an individual report on their child's ability, strengths, and weaknesses.

Because many educators attended this session hoping that McGraw-Hill was developing a direct writing assessment instrument to meet the needs of their school systems, Glick briefly discussed the publisher's new CAT Writing Assessment System, which offers interrelated options in the four areas of testing, prompts, scoring, and reports. For testing, it offers indirect assessment, direct assessment, or a combination. It offers two or three prompts per grade level. For scoring, the assessment system offers both professional evaluation in their Composition Evaluation Center in Monterey, California, and an Administration and Scoring Manual that can be used by local teachers to score papers. For reports, CAT can produce a variety of forms, including individual records, class roster reports, and frequency distributions.

The second speaker, Barbara Cole, Director of McGraw-Hill's Composition Evaluation Center, began her part of the session by echoing Glick's point that testers should always be concerned with the purpose of a test and by commenting on the fact that tests, scoring methods, and scores can all be used in inappropriate ways. Cole then described how the Composition Evaluation Center recruits, selects, and trains the evaluators for direct assessment of writing tests. The trainee program includes theory and training in holistic, analytic, and primary-trait scoring, as well as practice in the appropriate scoring of "validation packets" before trainees are allowed to score "live papers." Finally, table leaders constantly evaluate readers and their scoring to maintain the integrity of the program.

Cole noted that all prompts in the CAT system are two to four sentences long, with only one writing task specified in each. Tests take twenty-five to forty minutes, depending on the grade level being tested. She admitted the possibility of problems if prompts remain the same over the years, and she sees the need to develop and field-test equivalent prompts for the testing system.

---

ESSAY TEST TOPIC DEVELOPMENT

Speaker: Gordon Brossell, Florida State University
Introduction/Recorder: Bruce Holland, University of Akron, Ohio

Gordon Brossell posed the question, "What do we know about essay topic development for large-scale assessment?" at the outset of his session and proposed to answer it in light of recent research and practice. He first summarized the following maxim, principles, and guidelines relevant to making assignments for writing assessment, set forth in an ERIC report written by Leo Ruth and others in 1982.

Maxim: A well-framed assignment for writing assessment provides an occasion for writing in which students' powers of expression and communication are stimulated to their maximum abilities.

 Principles: 1) Subjects chosen should potentially be interesting to the teacher/writer.
2) Subjects chosen should potentially be interesting to the student/writer.
3) Data furnished should be open to presentation in several forms: e.g., statement, question, picture, cartoon.
4) The assignment should be meaningful within the student's experience.
5) The assignment should elicit a specific response and limit content or form or both.

Guidelines: Questions to consider in creating topics for writing assessment are:

1) What is the assignment's purpose (e.g., to ascertain placement, minimum competence, etc.)?
2) What will be done with the results?
3) What is the nature of the student population being assessed?
4) What is the most appropriate topic/subject for this group?

Brossell next identified three major situational variables in large-scale writing assessment:

1. Topic Variables: vocabulary and syntax; subject matter; the assumed or intended mode; rhetorical specification (i.e., information about purpose, audience, situation, etc.).

2. Writing Variables: topic interpretation; writer apprehension and anxiety.

3. Procedural Variables: instructions on the exam; time limit; scoring systems and readers; test environment.

Brossell discussed the effects of changes in vocabulary and syntax on writing assessment results. He noted that although common sense suggests such changes will make a difference to the writer, recent research has called this view into question. In Karen Greenberg's 1981 study at CUNY, variations in the cognitive demands and the experiential demands of a writing assignment were found to produce no significant difference in the essay scores. Gordon Brossell's own study, done with Barbara Ash, showed that neither wording changes nor subject changes had much effect on an essay's score. Brossell cautioned that he was not claiming that differences in subject are not important, but that we can't predict with certainty the degree and kind of change that will produce