marginal topics more accessible to a broader range of
students. In their closing remarks, the presenters emphasized the
most important considerations in topic selection. To the
extent possible, topics should be:
• clearly stated, using language familiar to the
audience
• appropriate to the population being tested
• reasonable for the allotted writing time
• suited to the purpose of the test
• accessible to the broadest range of students
• interesting to write about
• capable of producing writing across the range of
student abilities

Some participants were concerned about the limitation of
twenty minutes for writing the essay, particularly in light of
instructural emphasis on the importance of the prewriting,
drafting, composing, and revising processes. Although the
presenters acknowledged the potential difficulty of the time
constraint, they noted that it would be impractical in a
testing situation to provide students with the necessary time
and distance to simulate actual writing exercises. In a
separate project now nearing completion, ETS is analyzing the
influence of additional time on the quality of writing
produced by examinees taking the NJCBSPPT essay. Under
controlled conditions, students in that study were given
twenty minutes, thirty minutes, and twenty minutes with
ten minutes of prewriting; results of the investigation will
be available in late spring.

A separate question involved the comparability of scores
obtained using different topics. The presenters warned
against such direct comparisons, even when scores are
derived using trained readers and carefully selected
rangefinders. However, ETS is studying both the inter-
reader reliability across different scoring sessions involving
the same topic and the comparability of scores when the
same topic is repeated in separate test administrations.
These studies primarily focus on the effectiveness of reader
training.

TRAINING SESSION IN THE HOLISTIC
SCORING METHOD

Speaker: Rose Ann Morgan, Middlesex County
College, N.J.
Introducer/Recorder: Dennis Donahue, New Jersey
Institute of Technology

Rose Ann Morgan gave conferences attending this session a
hands-on introduction to holistic scoring. Participants had
an opportunity to read, score, and discuss examples of essays
written by students as part of the New Jersey College Basic
Skills Placement Test (NJCBSPPT). The session was a
scaled-down simulation of the training given to the readers,
usually high school and college teachers, who score these
essays.

Morgan explained that the essence of holistic reading and
scoring is the non-subjective appraisal of the work in
question. Readers are not asked to do anything but to assign
numbers to discourse. The numbers (1 to 6 on the
NJCBSPPT) represent different levels of achievement on the
test, but the significance of the numbers, the judgment
regarding which number represents passing or failing or
placement in one course or another, is to be made by others.
Such a judgment might be made, for example, by the person
at a college responsible for placing students in a course at an
appropriate level. Such people see the whole range and
distribution of actual scores. The function of a holistic
scoring training session is to make sure that all the readers
know and are able to employ the full scale and range of
scores for evaluating essays.

In an actual holistic scoring session, given just before a
reading, readers sit in groups of six or seven under the
guidance of a table reader, an experienced reader who has
participated in a preliminary holistic scoring session a day or
so prior to this general reading. The readers are given copies
of prescored sample essays, called rangefinders, and are asked
to learn—or re-learn if they are experienced readers—the scale
to be used throughout the reading. They are asked to rank
order these essays and then to assign each a score ranging
from 1 (lowest) to 6 (highest). There is at least one essay for
each of the six scores.

When all have finished scoring the rangefinders, the
readers are asked to announce their scores by a show of
hands. Readers can see instantly whether they are scoring
essays higher or lower than the table leaders and other readers
in the room. They then make any mental adjustments
necessary to bring their scoring into line with the scale
being used by the group. Additional rangefinders are scored
and discussed and then, when the chief reader is satisfied that
readers are using the agreed upon scale, "live" essays are
brought out and the actual scoring begins. At irregular
intervals during the session, additional rangefinders are
scored by the group to keep their idea of the scale sharp.
At the abbreviated training session given at the
conference, participants were given the same information
readers at an actual session get. They were told that
the writers had been given just twenty minutes to write an
expository response to a question relating to a task they had
not previously been able to perform but now could.
Participants were reminded that the writers' task was
complex, that a full answer might deal with the writer's past
and present and might give an explanation as to why the
writer now could perform that task. But Morgan asked
readers to read the essays supportively, and not to expect all
writers to respond to all three areas. Rather, the readers were
to score the essay on the basis of what had actually been
written. When Morgan was asked whether this was unfair,
she responded that it wasn't, that a given essay might cover
all three areas, not that it should. She added that a writer's
covering all three areas was no guarantee that the writing
was good, whereas an essay that had dealt with only one or
two areas before time elapsed might in fact be well written,
if incomplete.

A special method of training readers was then employed.
Participants were given just three rangefinder essays to begin
and were asked to rank them high, middle, and low. Then
they were given two additional papers and asked to interleave
them among the first three. Finally, they were given the
marginal topics more accessible to a broader range of students. In their closing remarks, the presenters emphasized the most important considerations in topic selection. To the extent possible, topics should be:

- clearly stated, using language familiar to the audience
- appropriate to the population being tested
- reasonable for the allotted writing time
- suited to the purpose of the test
- accessible to the broadest range of students
- interesting to write about
- capable of producing writing across the range of student abilities

Some participants were concerned about the limitation of twenty minutes for writing the essay, particularly in light of instructional emphasis on the importance of the prewriting, drafting, composing, and revising processes. Although the presenters acknowledged the potential difficulty of the time constraint, they noted that it would be impractical in a testing situation to provide students with the necessary time and distance to simulate actual writing exercises. In a separate project now nearing completion, ETS is analyzing the influence of additional time on the quality of writing produced by examinees taking the NJCBSPPT essay. Under controlled conditions, students in that study were given twenty minutes, thirty minutes, and twenty minutes with ten minutes of prewriting; results of the investigation will be available in late spring.

A separate question involved the comparability of scores obtained using different topics. The presenters warned against such direct comparisons, even when scores are derived using trained readers and carefully selected rangefinders. However, ETS is studying both the inter-reader reliability across different scoring sessions involving the same topic and the comparability of scores when the same topic is repeated in separate test administrations. These studies primarily focus on the effectiveness of reader training.

**TRAINING SESSION IN THE HOLISTIC SCORING METHOD**

**Speaker:** Rose Ann Morgan, Middlesex County College, N.J.

**Introductor/Recorder:** Dennis Donahue, New Jersey Institute of Technology

Rose Ann Morgan gave conference attendees attending this session a hands-on introduction to holistic scoring. Participants had an opportunity to read, score, and discuss examples of essays written by students as part of the New Jersey College Basic Skills Placement Test (NJCBSPPT). The session was a scaled-down simulation of the training given to the readers, usually high school and college teachers, who score these essays.

Morgan explained that the essence of holistic reading and scoring is the non-subjective appraisal of the work in question. Readers are not asked to do anything but to assign numbers to disclose. The numbers (1 to 6 on the NJCBSPPT) represent different levels of achievement on the test, but the significance of the numbers, the judgments regarding which number represents passing or failing or placement in one course or another, is to be made by others. Such a judgment might be made, for example, by the person at a college responsible for placing students in a course at an appropriate level. Such people see the whole range and distribution of actual scores. The function of a holistic scoring training session is to make sure that all the readers know and are able to employ the full scale and range of scores for evaluating essays.

In an actual holistic scoring session, given just before a reading, readers sit in groups of six or seven under the guidance of a table reader, an experienced reader who has participated in a preliminary holistic scoring session a day or so prior to this general reading. The readers are given copies of prescored sample essays, called rangefinders, and are asked to learn—or re-learn if they are experienced readers—the scale to be used throughout the reading. They are asked to rank order these essays and then to assign each a score ranging from 1 (lowest) to 6 (highest). There is at least one essay for each of the six scores.

When all have finished scoring the rangefinders, the readers are asked to announce their scores by a show of hands. Readers can see instantly whether they are scoring essays higher or lower than the table leaders and other readers in the room. They then make any mental adjustments necessary to bring their scoring into line with the scale being used by the group. Additional rangefinders are scored and discussed and then, when the chief reader is satisfied that readers are using the agreed upon scale, "live" essays are brought out and the actual scoring begins. At irregular intervals during the session, additional rangefinders are scored by the group to keep their idea of the scale sharp.

At the abbreviated training session given at the conference, participants were given the same information readers at an actual session get. They were told that the writers had been given just twenty minutes to write an expository response to a question relating to a task they had not previously been able to perform but now could. Participants were reminded that the writers' task was complex, that a full answer might deal with the writer's past and present and might give an explanation as to why the writer now could perform that task. But Morgan asked readers to read the essays supportively, and not to expect all writers to respond to all three areas. Rather, the readers were to score the essay on the basis of what had actually been written. When Morgan was asked whether this was unfair, she responded that it wasn't, that a given essay might cover all three areas, not that it should. She added that a writer's covering all three areas was no guarantee that the writing was good, whereas an essay that had dealt with only one or two areas before time elapsed might in fact be well written, if incomplete.

A special method of training readers was then employed. Participants were given just three rangefinder essays to begin and were asked to rank them high, middle, and low. Then they were given two additional papers and asked to interleave them among the first three. Finally, they were given the
last four rangefinders and asked to interleave them as they did the others. The readers were asked to be certain that each of the six scores was assigned to at least one of the essays. When the participants revealed their scores and the results were tabulated, they learned that they had achieved a fair amount of agreement among themselves and that their scores were reasonably close to scores given these essays at New Jersey Basic Skills scoring sessions. Agreement was most clear at the extremes of the scale, at the 6's and the 1's. Scores tended to be distributed more loosely in the middle of the scale. If time had permitted, participants here, like readers at an actual holistic reading, would have been given additional rangefinders to score so that they could sharpen their understanding of the middle range.

In a question and answer period that followed, Morgan discussed a number of issues, among them the procedures for determining the various score levels and for resolving debates about scoring a given essay.0

FROM HOLISTIC TO PROFILED SCORING OF SPECIFIC ACADEMIC WRITING

Speaker: Liz Hamp-Lyons, University of Michigan

Introducer/Recorder: John E. Naco, Community College of Philadelphia

Liz Hamp-Lyons presented conclusions from her research on the scoring procedures for the British Council Proficiency Test of the English Language Testing Service. She explained that the ELTS is charged with responsibility for testing overseas, non-native English speaking students who wish to enter Britain for post-graduate study. As such, its tests are roughly parallel to TOEFL instruments in the United States. The British Council Proficiency Test is composed of five components: reading, listening, study skills, writing, and oral interview (one-to-one). For the last three components, students choose 6 to be tested for skills within one of the several disciplines, including life sciences, medicine, social studies, technology, physical science, and general academic knowledge. The test is administered on demand to about 12,000 candidates per year in about ninety countries around the world. One person in each country is designated ELQ (English Language Qualified) and is responsible for scoring all tests administered in that country. In many countries, as few as five students per month take the test. For the writing component in every module, there are two readings to which candidates compose written responses. The first reading requires a "divergent" approach, and candidates must apply outside experience, opinion, judgment, or analysis to their reading when they write in response to this text. The second reading requires a text-based approach; candidates must restrict their written analysis to features in the reading itself. Scoring of the writing test component is criterion-referenced and scaled for nine levels (bands) of performance (from "non-writer" to "expert writer"). The test, originally administered in 1980, posed many problems at the outset. First, it was designed by English language professionals who lacked specific qualifications in test development and measurement. Secondly, persons designated ELQ in countries which administered the test would sometimes see only one candidate per month, making it impossible to collect a sufficient number of test performances for effective rating comparison. Finally, only one set of "benchmark papers" (a sample representing bands 8-3) was available for raters' reference. In addition, at this early stage of administration (1980-83) there was no test for reliability. It was not until 1984 that Hamp-Lyons was charged specifically with tightening up reliability and improving rating procedures.

In order to accomplish this, she started by looking at the descriptors across which tests were rated in the several bands; her purpose was to tease out the criteria that were implicit in the original descriptors. These descriptors were as follows: communicative quality (essentially holistic), organization (showing the extent to which logical structure carries the message), argumentation (effectiveness of argument), and linguistic accuracy and appropriateness (mechanics, syntax, etc.). She also discovered that some of the greatest problems candidates had in writing for the test were with length of response and plagiarism. Accompanying problems in these areas was the disturbing tendency of raters to assign a range of scores (for example, 4/5/6) on the writing test instead of confidently fixing a single score.

In answer to these immediate difficulties with the test, Hamp-Lyons wrote a training manual to accompany the test stating what the criteria meant, what the problems were, and how to deal with them. In order to effect these changes, she had to persuade the British Council to publish new short paragraph band descriptors for each of the nine levels. At the bottom of the revised, "profiled" scoring sheet is a box labeled "Final Band," for the average score on the writing. Admissions officers at British universities are interested in this score while the British Council looks at subscores to inform their decision whether to accept the foreign graduate student. The English Language Center at the student's university receives a copy of this sheet so that the university knows what the student's abilities are. Based on the student's performance across the criteria, the British Council can make a tentative decision that, for example, the student is a competent writer but probably a slow reader. In this case, the Council will fund the student to take an English "upper" course before starting at the university in October. At the University of Edinburgh, students were placed into sections according to the band at which they scored. Some students were at various bands upon admission to the university of their choice. Hamp-Lyons concluded by saying that she is attempting to introduce this "profiled" scoring procedure in placement testing at the University of Michigan, where she is currently Testing Coordinator.