(2) defining passing scores relative to the actual demonstrated performance of contemporary graduating seniors. Only those examinees who receive scores that are better than 30% of high school seniors are awarded the diploma. The job of the GED staff is rather to describe the skills and content knowledge that characterize the work of high school seniors than to prescribe levels of achievement.

The recent addition of an essay exam to the Writing Skills Test created questions about how to establish and maintain reliability. The GEDTS's first activity was to develop a scoring scale that would have the same criteria regardless of time or place. By administering an essay exam to thousands of high school seniors, sorting those essays into six stacks, and describing the characteristics of each stack, the Writing Committee of GEDTS was able to develop a holistic scoring scale that has been used successfully in hundreds of sites nationwide.

Auchter then reported on how GEDTS insures stability and reliability in the use of the scoring guide. A permanent GEDTS Writing Committee, consisting of practicing language arts professionals, selects the topics and the papers that are used in training, certifying, and monitoring site trainers and readers. The Writing Committee chooses and tests "expository" topics that do not require students to have any special knowledge or experience. The next step is for GEDTS to train and certify Chief Readers who are responsible for insuring that the GED scoring standards are applied uniformly. During the 2 1/2 day training, Chief Readers learn to overcome personal biases (e.g., responses to handwriting) that may influence scoring, and to use the language of the scoring guide alone to describe and evaluate papers. Sets of training papers contain a range of papers for each point, to illustrate the fact that there is no "perfect" paper for each point, but that there is typically a distribution of high, medium, and low papers. Training packets also include problematic papers (e.g., a paper written in the form of a rap song). Since the national average for high school essays scores is 3.25 and for GED scores is 2.7, training sets contain a disproportional number of 2, 3, and 4 papers. After working with training papers, GEDTS trainees are required to evaluate several sets of papers to determine whether or not they are currently certifiable as Chief Readers.

The same training and certifying procedure is carried out at the various decentralized testing sites, with the Chief Readers responsible for training and certifying readers. Auchter noted that language arts teachers trained through this process feel better about reaching writing and about using holistic grading in the classroom.

Further steps to insure score scale stability and reliability are site certification and monitoring. Each scoring site must demonstrate the ability to score essays in accord with the standards defined by the GED Testing Service. Essays used for site certification must receive at least 80 percent agreement in scoring among Writing Committee members. Although some sites may achieve high inter-reader reliability, a site cannot pass certification unless it achieves at least 90 percent agreement with GEDTS essay scores (or 85 percent for a provisional pass). Three procedures are used to monitor testing sites: (1) the Chief Reader does three readings of discrepant scores and records each time a reader is off the standard; (2) readers evaluate a set of "recalibration" papers at the beginning of each scoring session in order establish reliability for that day; and (3) GEDTS conducts site monitoring using the same procedures as are used in site certification.

TRAINING OF ESSAY READERS: A PROCESS FOR FACULTY AND CURRICULUM DEVELOPMENT

Speaker: Robert Christopher, Ramapo College, New Jersey
Introducer/Recorder: Mary Ellen Ashcroft, University of Minnesota

Robert Christopher emphasized the imperative of assessment, pointing out that assessment has always been intrinsic to the classroom experience, but it has now become extrinsic. He noted that many faculty fear writing assessment efforts because they represent an intrusion on their methods for evaluating students. He stated that faculty fears can be countered by several arguments: assessment helps students, it facilitates faculty and institutional research, and it is a professional activity.

Christopher went on to suggest ways of building faculty consensus for assessment. In a training readers, he suggested starting with a loyal, supportive core. This group's primary responsibility would be the development of an instrument for assessment, a task which should take six months to a year. He suggested that good readers are people who are task oriented, are good collaborators, are preferably not new faculty members (who might not have a sense of writing at the institution), and who work with "all deliberate speed." Good readers must not be "Matthew Arnolds" before whose standard everything fails. It is important, according to Christopher, that a large pool of readers be developed, so that a small loyal group will not wear out.

The next step, according to Christopher, is conducting a reading to build consensus. The initial reading should consist of 500 to 1000 essays, so that readers get a sense of the range of writing abilities of students at the institution. The reading must be conducted "blind" with each paper read and assessed twice. Essays are identified as "strong," "weak," and "in-between"; readers discuss each essay and slowly evolve into an interpretive community.

In terms of curricular implications, Christopher pointed out that placement assessment is easier to
accomplish and has been more fully studied than proficiency assessment. He also noted that assessment can be used for students to learn to talk about their writing in small groups and in conferences, so that students learn to to better readers and editors. Assessment can also be used to encourage collaborative or group teaching, said Christopher. As faculty members relinquish some control to group or collaborative situations in the assessment process, they learn from one another and share techniques and materials.

In answer to conference questions about developing the holistic process, Christopher suggested that the English Department provide a core of expert readers which should eventually grow to become interdisciplinary. He noted that in any two-day reading of essays, there is always the need for reliability checking ("Let's all read this essay and make sure we're on track"). Finally Christopher pointed out that students benefit from holistic essay assessment because their writing skills are evaluated by a team of teachers. This kind of assessment program, Christopher says, works on behalf of students.

**DISCREPANCIES IN HOLISTIC EVALUATION**

**Speakers:** Donald Daiker, Miami University, Ohio
Nedra Grogan, Miami University, Ohio

**Introduceor:** Sandra Flake, University of Minnesota

Donald Daiker presented the goals of the sessions: to share the conclusions and a tentative evaluation of his and Nedra Grogan's examination of discrepancies in holistic evaluation. Noting that discrepancies in holistic evaluation have been a problem from the beginning, he raised two questions: What accounts for discrepancies in holistic evaluation if the "quirky" reader is ruled out? And is there such a thing as a discrepant essay?

Daiker and Grogan sought to answer these questions using an annual holistic grading session for Miami University's Early English Composition Assessment Program (EECAP), a program in which 10,000 essays written by high school juniors in a controlled setting are evaluated for diagnostic purposes. The setting was one in which students, using a prompt, wrote for 35 minutes in a high school composition class. The time limitation was dictated by the constraints of a single class period. The goal of the holistic evaluation was essentially diagnostic, with a scoring scale of 1 to 6. Grades of 5 or 6 indicated clearly above average papers demonstrating strengths in all of the rating criteria. Grades of 3 or 4 indicated papers ranging from slightly below to slightly above average, with combined strengths and weaknesses in the criteria or under development. And grades 1 or 2 indicated clearly below average papers failing to demonstrate competence in several of the criteria, often because the paper was too short. A grade of 0 was used only for papers which were off the topic of the prompt. Evaluators gave each paper a single holistic rating, and additionally rated criteria in four categories (ideas, supporting details, unity and organization, and style).

The participating high school teachers (who were the evaluators) were trained through a process of rating and discussing sample papers, so that the rating criteria would be internalized. Participants in the session were then provided with the writing assignment or prompt, the scoring scale, the rating criteria, a rater questionnaire, and one of the papers.

To locate possible discrepant papers, Daiker looked for three-point gaps in scoring by two evaluators and gave such papers to both a third and fourth evaluator. If those evaluators also disagreed on the rating of the paper, he identified it as a potentially discrepant paper. Through this process, four potentially discrepant papers were identified, and those four papers were given to all 61 of the evaluators in a session at the end of the second weekend of evaluation. Participants in our session then read and evaluated one of the potentially discrepant papers, using a rater questionnaire, scoring scale, and rating criteria. The rating of the participants were tabulated: 1 person assigned the the paper a 6, 16 assigned a 5, 28 assigned a 4, and 4 assigned a 3.

Following the participant evaluation and some discussion, Grogan presented the result of the evaluation by 61 trained raters who rated the paper at the end of the second weekend of evaluation, with 25 of the raters (42.6%) giving an upper range (3-6) rating, 34 of the raters (55.8%) giving a middle range (3-4) rating, and 1 (1.6%), giving a lower range (1-2) rating.

Because of the clear division between the 5-6 and the 3-4 rating, Grogan and Daiker believe that the paper did qualify as a discrepant paper. Daiker reported that discussion following the rating by the trained evaluators suggested a correlation between the depth of emotional response to the paper and the highness of the score. Following some discussion about whether or not the paper was truly discrepant, a conference asked whether the problem was really caused by discrepant readers who could not be objective because of the depth of their emotional response. Daiker argued that reader objectivity was more complicated issue and further argued that precisely because the paper provokes a range of responses to the emotional content, it could be defined as a discrepant paper.

The implications of evaluating discrepant paper were then summarized by Grogan, who raised the issue of the role of holistic evaluation of a single essay that receives discrepant scores. She concluded that in such cases a single essay should not determine the fate of the writer, and that an appeals process clearly needs to be a significant part of a holistic evaluation program. Discussion throughout the session focused on some of the limitations of holistic evaluation of writing produced under a time constraint, on problems in establishing clear criteria and scales, and on problems of reader objectivity.