DISCREPANCIES IN HOLISTIC EVALUATION

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

Introducer/ Recorder: *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 26 of the raters (42.6%) giving an upper range (5-6) 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 conferee 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.