

PROBLEMS AND SOLUTIONS IN USING OPEN-ENDED PRIMARY TRAIT SCORING

Speaker: *Michael C. Flanigan, University of Oklahoma*

Introducer/

Recorder: *Chris Anson, University of Minnesota*

Michael Flanigan began by outlining his university's plan for five years of experimental research on the teaching and testing of writing. Much of this research will replicate published research studies, but original research will also be conducted. All of the studies will be controlled experimental studies so that the researchers can be fairly faithful to the original ones and can analyze any differences between the original and new research.

Flanigan discussed one study, already completed, in which his colleague David Mair and he combined the strategies of two studies by George Hillocks, an experimental study involving teaching extended definition using inquiry and models and a descriptive study dealing with "modes of instruction" (both of which Hillocks discusses in some detail in this book Research on Written Composition). In the replicated study at Oklahoma, all twenty classes consisted of university freshmen; for nine of the ten teachers it was their second semester of teaching, and approximately 500 students were involved.

Flanigan pointed out he chose Hillocks' studies because both dealt with significant areas in teaching and writing. Extended definition represents a kind of discourse that permeates almost all thinking and writing. The researchers believe that by replicating such an important study they could get inside the problems of the earlier research, and come to understand it better. The experimental extended definition study also used Hillock's open-ended primary trait scoring technique because the researchers wanted to learn to use and understand it better.

After reporting the findings from a small sampling of the data, Flanigan described some problems that he and his colleagues faced as they attempted to use Hillocks' open-ended primary trait scoring system and he discussed the modifications they made in it to obtain reliable results. He pointed out that with an open-ended primary trait scoring scale theoretically there is almost no limit to what students can score. Most scoring scales range from 1 to 6 (as in the holistic score for the ECT), 2 to 8 (as in CLEP), 1 to 5 (as in CORE scoring) and so forth. In open-ended primary trait scoring, the limit for a talented student is probably dictated by time and the variation and limitations imposed by the writing called for. In the papers scored in this study, the top score was 28.

The traits for which students could receive scores were: (1) properly putting an item in a class; (2) creating criteria for the class; (3) giving examples; and (4) providing contrastive examples to clarify and limit each criterion. Points were not given for differentiae as in Hillocks' original study; instead, class and differentiae

were combined (on the advice of Hillocks when the study was set up). Hillocks' scorers had had problems reaching agreement on this point. Students could receive 2 points for the class, 2 for each criterion, 2 for each example, and 2 for each contrastive example. Obviously the more criteria, examples and contrastive examples students could come up with, the higher their score. In initial training, scorers had problems staying close together in the higher ranges, so Flanigan modified his tolerance of acceptability by allowing scores in the range 1 to 10 to differ by 1 point, 11 to 20 to differ by 2 points, and 21 up to differ by 3 points. Scores within that range were averaged; scores that did not meet acceptable standards were read by a third reader. If the third reading fell within range of either of the other two readers, then those scores were averaged. If there still was no agreement, a fourth and fifth reader scored the paper, and the paper and the range of scores were given to the researchers and a score was determined. For example, one paper was scored 6 and 8; a third reader gave it 10; the fourth reader gave it 9, and the fifth reader gave it 7. Its final average was an 8. Only seven papers required the fourth and fifth reader. Often, readers had problems keeping clearly in mind the kinds of criteria the writers were developing. To simplify the process, any one clear criterion could be accompanied by a number of examples and contrastive examples. If no criterion was given, only one example could be counted. If an undeveloped example or string of general examples was given, a score of 1 was given.

Flanigan concluded that open-ended primary-trait scoring offers real promise, for it allows for a kind of differentiation that closed, limited systems do not. However, researchers who use the system will probably have to modify it to get consistent, reliable scores. They will also have to plan their research so that the traits they are describing and scoring are clear, well-defined, and fully conceptualized by their scores. The session ended with the speaker giving participants six papers that had been scored by three readers and leading participants through a guided scoring session.