MAINTAINING SCORING STANDARDS IN LARGE-SCALE WRITING ASSESSMENT: MEDIATING BETWEEN TEACHERS AND TRAINED READERS

Gail Lynn Goldberg, Maryland State Department of Education

This presentation focused on the Maryland Writing Test (MWT). A hallmark of this test is the reliance upon the Scoring Committee (SC), a corps of classroom teachers and writing specialists, for rangefinders. This practice facilitates the correspondence between writing instruction and assessment, gives the Maryland educational community a needed sense of ownership of test results, and removes any sense of the "tyranny of testers." The foremost purpose of the SC is to maintain scoring standards; toward this end committee turnover is minimized, and the group participates in annual recalibration on the standard-setting year (1984) before scoring nearly 720 responses from that year's administration. MWT training papers are then selected from those receiving SC consensus (> 70% agreement). Score stability would appear to require consistency not only in personnel and procedures, but in the development and application of SC data as well.

The MWT requires that students respond to each of two prompts, one narrative and one explanatory. Two trained readers score all responses, including those prescored by SC. The MWT is scored using a modified holistic procedure based on a four-point, domain-specific rubric, and adjacent agreement is acceptable. 1987 marked the earliest attempt to investigate discrepancies between SC and trained-reader-scores. Since then, a detailed tracking system has

been implemented which permits comparison of scores given by "expert readers" (MSDE writing assessment specialists and contractor's senior staff), SC, and trained readers.

In 1989, expert readers reexamined all responses on which there was a score discrepancy. These included papers used in training, those with an SC consensus score not used in training, and papers coded "do not use" (DNU) either because SC consensus was not reached or SC and expert readers disagreed. The purpose of this review was not to investigate score reliability; rather, it was to identify scoring trends exhibited by both groups and to consider how to explain and attend to the score deviation.

The 1989 score discrepancy data suggest that direct writing assessment programs exchange the linearity of the usual rangefinding and application process for a more complex reintegration of information on scores and scoring behaviors. Results of this analysis of SC/reader discrepancies were reported to SC and built into their recalibration and training process for an upcoming summer administration of the MWT. SC reviewed salient examples so members could ascertain text features that contributed to discrepant evaluations and could hone scoring lines further. Stability of scoring standards will be further enhanced in future years by accelerating the descriptive analysis process to allow feedback to readers during scoring, with use of papers with discrepant scores as supplementary training materials. Stability ought not be synonymous with procedural "status quo"; interchange among all groups involved in the maintenance of scoring standards, with concomitant changes in scoring practice and procedures, may be the best assurance of stable standards.