MAINTAINING SCORING STANDARDS IN LARGE-SCALE ASSESSMENT: A HERACLITEAN PERSPECTIVE

Belita Gordon, The University of Georgia

This session opened with questions about a revered assessment tenet: namely, that scoring standards remain constant over time. While score stability is necessary to ensure fairness, the pragmatics of large-scale testing make consistency difficult to attain. Instruction improves the quality of student writing, different contractors rate the same state's papers from year to year, prompts of inherently different difficulty are administered, and scoring guides and training procedures are refined. We respond with archival data and statistical equating procedures that anchor us to our past, obligating us to repeat not only our successes, but also our errors. Stability may be paramount, but this does not negate the need for change. The Basic Skills Writing Test (BSWT) is a case study of the benefits of change.

The BSWT is administered to Georgia's tenth graders as one of the requirements for a high school diploma. In the development stages, teachers expressed the desire for a diagnostic test. Consequently, they rejected a single holistic

score in favor of five domain scores. The domains are holistically scored on a four-point scale. Two are weighted (the Content/Organization score is multiplied by three and the Style score by two), while the scores in Sentence Formation, Usage, and Mechanics are taken at face value. Students receive a total score and diagnostic statements for each domain. Over the two pilot years, the metric was changed from a two- to a four-point scale, as the two-point scale did not reflect the range of writing produced, and the Content and Organization domains were combined, as raters were unable to differentiate between the two.

An evaluation of the ratings on 1,866 papers, following the first operational year, strongly suggested that the domain subscores were too closely related to provide the desired diagnostic information. For the second operational year, the training procedure was revised and a comparative study implemented. Scoring accuracy was monitored, by domain, on the basis of agreement with prescored papers embedded within the packet of 80,000 "live" papers. The 1,866 anchor papers were scored again.

The analysis of the anchor papers revealed greater discrimination between domains. The highest correlation (between Content/ Organization and Style) dropped from .72 to .49 while the lowest (between Content/Organization and Usage) dropped from .49 to .40. An increase in the absolute mean difference for each comparison further reinforced evidence of greater differentiation. The frequency with which raters assign the same score in all domains has dropped from 42% to 25% on the first rating, from 37% to 22% on the second rating, and from 19% to 7% on both ratings. But while all three analyses suggest greater discrimination, they do not answer important questions. Does the test measure different subskills? Do the subscores provide instructionally useful information? Is writing a "unitary trait," and if not, how much discrimination is necessary before different attributes (domains) can be reported? Participants discussed these and related questions.