

# Packet for Designated Assessment Leaders in Departments George Mason University

*Materials for the normed scoring session developed by  
Ruth Fischer, Chris Thaiss, and Terry Zawacki*

## **Rationale:**

**Precedent:** The procedure we'll be practicing today is a modified version of the process that the Educational Testing Service (ETS) has been using for many years to calibrate standards for essays submitted by students attempting to gain credit for Advanced Placement in colleges and universities. The procedure enables a large group of evaluators to achieve consensus on standards efficiently and with a high rate of co-rater consistency. At Mason, we in English have been using this procedure for about ten years in the grading of student essays submitted for English 101 proficiency credit.

**Features:** (1) This is a **norm-referenced**, rather than **criterion-referenced**, procedure; i.e., the procedure depends on comparison of one sample essay against others in a group, rather than in matching of a given essay against a previously-established group of criteria. Nevertheless, one goal and benefit of the procedure is that it allows a group of raters to establish criteria that can then be applied in courses and in other evaluative situations. That is our purpose today.

(2) This is a **holistic** procedure rather than a **primary-trait** procedure. That is, the comparative ratings you will be asked to give will be based first on your overall (holistic) impression of the quality of the essays. Only after your impressionistic rating has been given will the group be asked to identify criteria (traits) that have determined your ratings. These traits can then be applied to subsequent readings.

## **Benefits:**

- 1. Buy-in.** The criteria defined in the session are those determined by the group, not imposed on them. Because the group has arrived at these criteria in discussion, there is a high level of mutual understanding of terms.
- 2. Validity.** Because criteria follow from the reading of actual student samples, they are relevant to the sample, hence meet the test of validity for writing assessment (White). (Accurate writing assessment has been bedeviled for years by the use of invalid instruments; e.g., the use of multiple-choice grammar quizzes to measure writing proficiency.)
- 3. Applicability across Disciplines.** For our purposes as a cross-departmental, cross-disciplinary group, the procedure is therefore one that can be used productively with other groups of faculty attempting to define criteria for student products that may differ greatly from discipline to discipline

4. **Reliability.** If readers in a session go through the procedure for several sets of comparisons, they achieve a high degree of “interrater reliability.” We are modifying the ETS procedure by limiting the number of comparative ratings, in the interest of time; however, our experience in the English proficiency readings has been that even as few as three iterations of comparison allow us to achieve a satisfactory level of consistency among readers.
5. **Efficiency.** We have found, as has ETS, that this holistic procedure can enable faculty to achieve a valid, reliable, consensual standard for assessment of a group of student papers in an hour or ninety minutes. Because faculty have gone through an inductive process to determine these criteria, those faculty are more likely to understand the terms of the criteria and apply them effectively in future contexts than they would had they been given a pre-determined, undefined list.

**Modified Applications:** We are not asking that faculty apply these methods in their own departmental contexts exactly as we will be practicing the procedure today. There are no doubt variations that will produce as good results and that better fit the context of the department. During the last part of the workshop we will be discussing options that you suggest. Nevertheless, we do strongly recommend some version of this procedure, so that departments can achieve consensus among faculty, as well as the validity and reliability of scoring that the method promotes.

For example, if a department uses this procedure toward determining how to measure the writing competency of its majors (part of the general education “synthesis” requirement), it might bring a group of faculty together to grade holistically samples of a typical paper in that course. Criteria defined in the session might then be used toward writing the assignments for that project in a subsequent course. Alternatively, a department might wish to measure competency by evaluating portfolios of work from more than one course; in that case, a faculty group could read holistically several portfolios, to help the faculty determine both (1) its criteria for competence and (2) the most appropriate contents of future portfolios.

White, Edward M. “Holistic Scoring: past Triumphs and Future Challenges.” In *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations*. Ed. Michael M. Williamson and Brian Huot. Cresskill, NJ: Hampton, 1993, 79-108.