MAINTAINING SCORE-SCALE STABILITY AND READING RELIABILITY IN DECENTRALIZED LARGE-SCALE ESSAY SCORING PROGRAMS

Wayne Patience and Joan Auchter, GED Testing Service

A central aim in any writing assessment program is to ensure score-scale stability and reading reliability across essay scoring sessions. In large-scale writing assessment programs, the essay scoring is often necessarily decentralized; that is, the scoring sites vary in location, personnel, frequency of scoring sessions, etc. In such large-scale essay scoring programs, it is essential to construct training, certifying, and monitoring procedures that are designed to establish scoring standards and ensure the consistent and reliable application of these scoring standards across scoring sites and scoring sessions.

The revised Tests of General Educational Development (GED) that were introduced in 1988 include a direct measure of writing as part of the Writing Skills Test. Because the Writing Skills Test scores, along with the scores on the other four subject-area tests in the battery, determine whether an examinee receives a high school equivalency credential, the reliability of the weighted composite score based on a combination of the essay and multiple-choice parts of the Writing Skills Test is of major importance.

This presentation described the methods of establishing and monitoring score-scale stability and reading reliability and reported results showing how well these methods succeeded in the first year of the GED Testing Program’s decentralized essay scoring operation. Several GED Testing Service scoring sites participated in monitoring two types of essay-scoring. Random monitoring was conducted by randomly sampling and reading essays that are scored at the decentralized sites and routinely sent to the GED Testing Service. Systematic monitoring was conducted by requiring sites to read and return previously scored and selected essay sets provided by the GED Testing Service.

Results showed that the procedure for establishing and maintaining score-scale stability are effective. In the minority of sites wherein some scale drift or inconsistency among readers is present, monitoring reports sent to site chief readers provide diagnostic information that shows the nature of the problem and specifies corrective action.

WRITING ASSESSMENT IN QUEBEC: THE STATE OF THE (LANGUAGE) ARTS

Tom McKendry, Marianopolis College, Montreal Allan Patenaude and Beverly Steele, Quebec Ministry of Education

This session began with a description of the integrated secondary Language Arts Curriculum prescribed in 1982 by the Quebec Ministry of Education (M.E.Q.). For the first time, teachers throughout the province faced a rationale and methodology of teaching English predicated on the principles that language serves as an agent of communication and conceptualization. In 1990, the common provincial exam for graduating students will be based on this new curriculum.

This common exam has served as a catalyst to the province wide discussion of several major issues: Should examination questions be linked directly to specific works? Should talk occur during the examination? Should rough drafts and outlines be graded as part of the examination? Each of these issues is currently being addressed by the M.E.Q. through in-service training and the development of evaluation models.

Over the past several years we have used both college English grades and midterm ratings of writing ability by college English teachers as standards against which to measure the effectiveness of various methods of placement in writing courses. Although the research was originally intended to compare the relative reliability of a holistically scored writing sample and a multiple-choice test, statistics on secondary school records were also gathered. Recent data analyzes indicate that the correlations between the secondary school grades and college standards have risen dramatically.

Although it is premature to conclude that the improved correlations are the result of the new secondary school curriculum, the data encourage further testing of this hypothesis.
A central aim in any writing assessment program is to ensure score-scale stability and reading reliability across essay scoring sessions. In large-scale writing assessment programs, the essay scoring is often necessarily decentralized; that is, the scoring sites vary in location, personnel, frequency of scoring sessions, etc. In such large-scale essay scoring programs, it is essential to construct training, certifying, and monitoring procedures that are designed to establish scoring standards and ensure the consistent and reliable application of these scoring standards across scoring sites and scoring sessions.

The revised Tests of General Educational Development (GED) that were introduced in 1988 include a direct measure of writing as part of the Writing Skills Test. Because the Writing Skills Test scores, along with the scores on the other four subject-area tests in the battery, determine whether an examinee receives a high school equivalency credential, the reliability of the weighted composite score based on a combination of the essay and multiple-choice parts of the Writing Skills Test is of major importance.

This presentation described the methods of establishing and monitoring score-scale stability and reading reliability and reported results showing how well these methods succeeded in the first year of the GED Testing Program’s decentralized essay scoring operation. Several GED Testing Service scoring sites participated in monitoring two types of essay-scoring. Random monitoring was conducted by randomly sampling and reading essays that are scored at the decentralized sites and routinely sent to the GED Testing Service. Systematic monitoring was conducted by requiring sites to read and return previously scored and selected essay sets provided by the GED Testing Service.

Results showed that the procedure for establishing and maintaining score-scale stability are effective. In the minority of sites