LARGE-SCALE WRITING ASSESSMENT IN CANADA: ISSUES AND PRACTICES

Speaker: Peter Evans, *Ontario Institute for Studies in Education (OISE) Ottawa Valley Centre*

Introducer/Recorder: Dale Oscarson, *Palo Alto Unified School District*

In his initial remarks, Peter Evans suggested that large-scale writing projects in Canada, specifically in Manitoba, British Columbia, New Brunswick, and Ontario, have led many participants to recognize the need for including writing to assure valid assessment of student performance in English. Evans noted that "The evaluation of a student is most reliable when based on several different examples of the student's work and writing over a period of time..." For instance, the Ontario Academic Credit English guidelines suggest that thirty percent of the final grade be based on the writing folder. Although Evans said he does not know the extent to which writing as a process has influenced classrooms across Canada, he cited instances of efforts in that direction involving many teachers and consultants, and he noted that the Canadian Council of Teachers of English evaluation policy states that:

> As far as possible, assessment should employ direct rather than indirect measures of achievement, and it may also consider process as well as product.

Secondly, Evans spoke to the topic of scorer behavior in writing evaluation. The first study to which he referred—the 1977-1980 OISE Writing Evaluation Project in Ontario for Grades 7 to 12, based its findings on 110 papers scored holistically by at least five scorers. The researchers took a close look at the specific writing features that affect reader behavior. These include content and organization, error frequency, essay length, and spelling. These features seem to correlate most consistently with scores across grade levels and modes. Analytic scoring suggests that teachers attend more to content and organization than to errors when arriving at holistic scores.

Evans devoted the final portion of his presentation to his study, "Sources of Rater Disagreement in Holistic Scoring," in which he analyzed holistic scores assigned to 640 essays: half narrative in response to a single stimulus, and half argumentative exposition in response to a single stimulus. Evans and his colleague, Philip Nagy, chose fifty essays on which scorers disagreed and fifty random essays from the second and fourth quintiles. Experienced English teachers received seventy-five essays for scoring, twenty-five from the "disagreement" category, twenty-five second quintile papers and twenty-five fourth quintile essays. These teachers holistically scored each essay. A word count for length and for error frequency per one-hundred words were provided by another group of scorers utilizing analytical scoring and error counting. A third group looked at expository and persuasive strategies through "framework retrieval," a complex analysis similar to "primary trait" scoring.

Based upon indirect analysis, the following conclusions were reached:

- High error frequency alone does not appear to be a source of scorer disagreement.
- For the argumentative essay there is no relationship between "essay quality" and frequency of errors. For the narrative essay there is a slight relationship (with higher quality essays having fewer errors).
- Length, in the expository essays, is clearly not a factor in scorer splits; for narrative essays, length does not influence scorer judgments. Brevity, at the other extreme, is irrelevant for the narratives under scrutiny, and for exposition, as one might expect, brevity and low quality scores tended to be associated.
- While expository essays high on quality of argument were distributed through the whole set to much the same degree as the subset under scrutiny, when "essay quality" and "quality of argument" scores were compared for the subset, there was frequent disagreement among ratings. It is reasonable therefore to hypothesize that disagreements about the relative importance of style and logic is a source of rater disagreement.

Several conclusions of interest to both large scale assessment and classroom teacher assessment might be considered as a result of the study:

1) Because a writer may offer one of a number of different types of narratives in response to a stimulus, scorers need to consider which of these treatments are legitimate responses.

2) Writers may be prized for broad, popular but unsupported sentiments as opposed to supported, unpopular ones.

3) Stylistic shifts should not be considered in themselves but should be considered in relation to writer purpose and audience.

4) Excessive formality should be regarded as inappropriate to audience or topic and not as a try at immortal prose.

5) Where a writer is unable to sustain a fine beginning into a finely wrought essay, scorers should look closely at the purpose for the particular assessment and grade accordingly.

6) A writing stimulus should be clear. As Evans concluded, "In writing assessment, the decision concerning modes and specific stimuli must be thought through very carefully and then pre-tested before large-scale use."